65
FEB 73
D-5
2123

NAVAL RESEARCH LOGISTICS QUARTERLY

SEPTEMBER 1972
VOL. 19, NO. 3

OFFICE OF NAVAL RESEARCH

NAVSO P-1278

407-B

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORS

H. E. Eccles
Rear Admiral, USN (Retired)

O. Morgenstern
New York University

F. D. Rigby
Texas Tech. University

D. M. Gilford
U.S. Office of Education

B. J. McDonald
Office of Naval Research

S. M. Selig
Managing Editor
Office of Naval Research
Arlington, Va. 22217

ASSOCIATE EDITORS

R. Bellman, RAND Corporation
J. C. Busby, Jr., Captain, SC, USN (Retired)
W. W. Cooper, Carnegie Mellon University
J. G. Dean, Captain, SC, USN
G. Dyer, Vice Admiral, USN (Retired)
P. L. Folsom, Captain, USN (Retired)
M. A. Geisler, RAND Corporation
A. J. Hoffman, International Business Machines Corporation
H. P. Jones, Commander, SC, USN (Retired)
S. Karlin, Stanford University
H. W. Kuhn, Princeton University
J. Laderman, Office of Naval Research
R. J. Lundegard, Office of Naval Research
W. H. Marlow, The George Washington University
R. E. McShane, Vice Admiral, USN (Retired)
W. F. Millson, Captain, SC, USN
H. D. Moore, Captain, SC, USN (Retired)
M. I. Rosenberg, Captain, USN (Retired)
D. Rosenblatt, National Bureau of Standards
J. V. Rosapepe, Commander, SC, USN (Retired)
T. L. Saaty, University of Pennsylvania
E. K. Scofield, Captain, SC, USN (Retired)
M. W. Shelly, University of Kansas
J. R. Simpson, Office of Naval Research
J. S. Skoczylas, Colonel, USMC
S. R. Smith, Naval Research Laboratory
H. Solomon, The George Washington University
I. Stakgold, Northwestern University
E. D. Stanley, Jr., Rear Admiral, USN (Retired)
C. Stein, Jr., Captain, SC, USN (Retired)
R. M. Thrall, Rice University
T. C. Varley, Office of Naval Research
J. F. Tynan, Commander, SC, USN (Retired)
J. D. Wilkes, Department of Defense OASD (ISA)

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Information for Contributors is indicated on inside back cover.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C. 20402. Subscription Price: $5.50 a year in the U.S. and Canada, $7.00 elsewhere. Cost of individual issues may be obtained from the Superintendent of Documents.

The views and opinions expressed in this quarterly are those of the authors and not necessarily those of the Office of Naval Research.

Issuance of this periodical approved in accordance with Department of the Navy Publications and Printing Regulations, NAVEXOS P-35

Permission has been granted to use the copyrighted material appearing in this publication.

MATHEMATICAL CONTROL THEORY SOLUTION OF AN INTERACTIVE ACCOUNTING FLOWS MODEL*

Yuji Ijiri

and

Gerald L. Thompson

Graduate School of Industrial Administration
Carnegie-Mellon University

ABSTRACT

In this paper we have applied the mathematical control theory to the accounting network flows, where the flow rates are constrained by linear inequalities. The optimal control policy is of the "generalized bang-bang" variety which is obtained by solving at each instant in time a linear programming problem whose objective function parameters are determined by the "switching function" which is derived from the Hamiltonian function. The interpretation of the adjoint variables of the control problem and the dual evaluators of the linear programming problem demonstrates an interesting interaction of the cross section phase of the problem, which is characterized by linear programming, and the dynamic phase of the problem, which is characterized by control theory.

1. INTRODUCTION

This paper is a sequel to our previous paper [1] which dealt with a wheat trading model in which only one product is to be bought and sold subject to maximum purchase and sales limits, when the price variation over time is known with certainty.

In this paper, we shall extend the applications of mathematical control theory to multiple product cases where more than one product is to be bought (produced) and sold under price certainty, subject to various kinds of linear constraints on purchases (production) and sales. We will use notation similar to that used in [1].

Our results are as follows: The adjoint variables have interpretations similar to those in the previous paper [1]. The optimal control policy is of the "generalized bang-bang" variety which is obtained by solving at each instant in time a linear programming problem whose objective function parameters are determined by the "switching functions" of the control problem. The optimal policy then "bangs" from one extreme point of this linear programming problem to another as the switching functions change over time. A simple example is worked out to show how the optimal policy changes as prices change over time.

2. THE CONTINUOUS JOINT TRADING MODEL

We shall describe this model in terms of a trading situation, but it can easily be reinterpreted in

*This report was prepared as part of the activities of the Management Science Research Group, Carnegie-Mellon University, under Contract NONR 760(24), NR 047–048 with the Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

terms of a model in which there is the joint production and sale of two different products.

Let us consider a firm that engages in trading two kinds of commodities, wheat and rice. The firm's assets are of only three kinds, cash, wheat, and rice, whose balances at time t are represented by $c(t)$, $w(t)$, and $v(t)$, respectively. These stocks are measured as follows: $c(t)$ in dollars, and $w(t)$ and $v(t)$ in bushels or in some other physical units. The price of wheat and the price of rice from time 0 to time T are assumed to be known in advance with certainty and are given by functions $p_1(t)$ and $p_2(t)$, respectively. It is also assumed that for each commodity there is only one price at any point in time which is applicable both for buying and selling the commodity. We shall let $P_1 = p_1(T)$ and $P_2 = p_2(T)$ be the terminal prices of wheat and rice, respectively. We assume that the initial assets $C_0 = c(0)$, $W_0 = w(0)$, and $V_0 = v(0)$ are known constants.

The firm's objective is to buy and sell wheat and rice during the time period 0 to T so as to maximize the total value of its assets at time T. Hence, it always values its current inventories with the terminal prices in making trading decisions. Thus the total value of the assets of the firm at time t, denoted by $x(t)$, is given by

$$x(t) = c(t) + P_1 w(t) + P_2 v(t). \tag{1}$$

We denote the rates of buying wheat and rice at time t by $m_{1p}(t)$ and $m_{2p}(t)$, respectively, and the rates of selling wheat and rice at time t by $m_{1s}(t)$ and $m_{2s}(t)$, respectively. They are all expressed in physical units of wheat and rice. We shall impose the following constraints on these variables which are assumed to reflect the buying and selling environments of the firm's operations.

$$\begin{aligned} A_p m_p &\leq b_p \\ A_s m_s &\leq b_s \\ m_p, m_s &\geq 0, \end{aligned} \tag{2}$$

where the matrices A_p $(k \times 2)$, A_s $(n \times 2)$, b_p $(k \times 1)$, and b_s $(n \times 1)$ are given constant matrices, and m_p is the column vector with components m_{1p} and m_{2p}, and m_s is the column vector with components m_{1s} and m_{2s}.

Sometimes there are additional constraints that connect buying and selling activities of the firm of the form,

$$D_p m_p + D_s m_s \leq b_d, \tag{3}$$

where the matrices D_p $(h \times 2)$, D_s $(h \times 2)$, and b_d $(h \times 1)$ are given constant matrices, and D_p and D_s each containing at least one nonzero constant in each row.

The only assumption that we shall make concerning the convex set defined by the constraints (2) and (3) is that there is *at least one* feasible vector satisfying them. Thus the constraint set may be unbounded. As we shall see, this may lead to a maximizing linear programming problem with unbounded solution without affecting our main results. Of course, in any reasonable application the solution to

such a maximizing problem will turn out to be bounded since unbounded solutions cannot be realized in practice.

Many other kinds of constraints can be imagined. For instance, constraints (such as those involving liquidity or cash flow) over time, or constraints involving both state and control variables. However, we shall restrict this paper to the consideration of problems that involve only the constraints in the form of (2) and (3) and that there is at least one feasible vector satisfying them.

As long as the constraints (2) and (3) are satisfied, the firm is permitted to be either long or short in any or all of the three asset types, cash, wheat, and rice. A long position is indicated by a positive stock of an asset and a short position by a negative stock.

There are four factors that change the cash balance. One is the interest received on cash (or paid on a loan) balance at a given constant rate R per unit period, which is to be accrued continuously and is to be received (or paid) in cash. That is, the interest on a cash or loan balance is compounded continuously at interest rate R. The second factor is the inventory holding costs $h_1(w)$ for wheat and $h_2(v)$ for rice that must be paid in cash on inventories of sizes $w(t)$ and $v(t)$ at time t. The functions h_1 and h_2 are piecewise differentiable functions of w and v, respectively. They are increasing functions of w and v for their nonnegative values and decreasing functions for the nonpositive values. Thus we interpret the inventory holding cost to be the cost of being short in wheat or rice when $w<0$ or $v<0$, which increases as $|w|$ or $|v|$ increases. The third factor changing the cash balance is the cost of purchasing wheat and/or rice, $p_1(t)m_{1p}(t)+P_2(t)m_{2p}(t)$ at time t. The fourth factor is the revenue from sales of wheat and rice, $p_1(t)m_{1s}(t)+p_2(t)m_{2s}(t)$. Thus, the differential equation that expresses the manner in which the stock of cash changes is

$$\dot{c}(t)=Rc(t)-h_1(w)-h_2(v)-p_1(t)m_{1p}(t)-p_2(t)m_{2p}(t)+p_1(t)m_{1s}(t)+p_2(t)m_{2s}(t). \tag{4}$$

There are only two factors that affect the stock of wheat and rice, namely the purchase and sales quantities. Thus, the differential equations for w and v are

$$\dot{w}(t)=m_{1p}(t)-m_{1s}(t) \tag{5}$$

and

$$\dot{v}(t)=m_{2p}(t)-m_{2s}(t). \tag{6}$$

In order to simplify notation, we shall drop (t) from all expressions with the understanding that lower case letters (except b and e) represent functions of time unless other independent variables are specified (e.g., $h(w)$), and that all upper case letters represent constants.

To put the problem into standard Mayer form, we differentiate (1) obtaining:

$$\dot{x}=\dot{c}+P_1\dot{w}+p_2\dot{v}=Rc-h_1(w)-h_2(v)-p_1m_{1p}-p_2m_{2p}+p_1m_{1s}+p_2m_{2s}+P_1m_{1p}-P_1m_{1s}+P_2m_{2p}-P_2m_{2s}. \tag{7}$$

The latter expression was obtained by using (4)–(6).

The firm's objective function is now expressed by

$$\text{Max } x(T)=c(T)+P_1w(T)+P_2v(T) \tag{8}$$

subject to the differential equations (4)–(7) and the constraints (2) and (3).

In order to solve the problem, we introduce adjoint variables $\Psi(t)$ for (4), $\phi(t)$ for (5), $\rho(t)$ for (6) and the constant function -1 for (7). Then the Hamiltonian function is given by

$$\begin{aligned}(9)\qquad \mathscr{H}(t) &= \Psi(Rc - h_1(w) - h_2(v) - p_1m_{1p} - p_2m_{2p} + p_1m_{1s} + p_2m_{2s}) + \phi(m_{1p} - m_{1s}) \\ &\quad + \rho(m_{2p} - m_{2s}) - (Rc - h_1(w) - h_2(v) - p_1m_{1p} - p_2m_{2p} + p_1m_{1s} + p_2m_{2s} \\ &\quad + P_1m_{1p} - P_1m_{1s} + P_2m_{2p} - P_2m_{2s}) \\ &= (\Psi - 1)(Rc - h_1(w) - h_2(v)) - \lambda_1(m_{1p} - m_{1s}) - \lambda_2(m_{2p} - m_{2s}),\end{aligned}$$

where λ_1 and λ_2 are the *switching functions* defined as

$$(10)\qquad \lambda_1 = P_1 + p_1(\Psi - 1) - \phi$$

and

$$(11)\qquad \lambda_2 = P_2 + p_2(\Psi - 1) - \rho.$$

The determination of Ψ, ϕ, and their interpretations together with those of $\mathscr{H}$, λ_1 and λ_2 are given in Section 4. They are entirely analogous to the corresponding functions in Reference [1].

The maximum principle [2], provides as a necessary condition for an optimal policy that the control variables m should be chosen so as to minimize $\mathscr{H}$, which implies the maximization of the criterion function

$$(12)\qquad z = \lambda_1(m_{1p} - m_{1s}) + \lambda_2(m_{2p} - m_{2s}) = \lambda(m_p - m_s),$$

subject to the control variable constraints (2) and (3). In other words, at each point of time, the optimal policy involves the solution of an ordinary linear programming problem. The coefficients of the objective function in (12) are the switching functions λ_1 and λ_2. Since these are functions of time, it follows that what is involved is a parametric solution of this linear programming problem. We summarize our results in the following theorem.

THEOREM: If the constraint set defined by (2) and (3) is nonempty, then the optimal control variables m are determined by solving at each time t the linear programming problem:

$$\text{Max } \lambda(m_p - m_s).$$

Subject to

$$(13)\qquad \begin{aligned} A_pm_p &\leq b_p, \\ A_sm_s &\leq b_s \\ D_pm_p + D_sm_s &\leq b_d \\ m_p, m_s &\geq 0, \end{aligned}$$

where $\lambda = (\lambda_1, \lambda_2)$ is given by (10) and (11).

PROOF: The calculations above show that necessary conditions for optimal control are that the control variables should solve at each point in time the linear programming problem (13). We must still show that these necessary conditions are sufficient. The sufficiency follows immediately, however, by observing that the constraints of the linear programming problem can be assumed (by well-known results) to be nondegenerate so that the solution to the linear programming problem is unique, except for the finite number of values of $\mathcal{A}_1$ and $\mathcal{A}_2$. The example in the next section illustrates the result.

It should be remarked that, if the solution to (13) is infinite, as may well happen for certain values of the switching function in the case that the constraint set is unbounded, our conclusion is still valid even though this gives an infinite value to the Hamiltonian. It may also happen that both m_p and m_s become infinite while the quantity $m_p - m_s$ remains finite, so that the Hamiltonian remains finite. This case also causes no difficulty. Obviously neither of these possibilities will arise in a problem derived from an actual application since solutions involving infinite rates of buying cannot realistically be carried out over any finite length of time.

There are two special cases worthy of mention and further detail. The first case is that of *separable constraints* which occurs when constraints of the form (3) are absent. Here the linear programming problem (13) breaks up into two smaller problems and hence it may be considerably easier to solve. The second case is that of *coordinated control* which occurs whenever $m_p^{tr} \cdot m_s = 0$. In other words, a coordinated control problem is one in which there is never a reason to simultaneously buy and sell at a given time. Again, in realistic applications this is most likely to be the case because of the normal workings of the market.

3. AN EXAMPLE

In this section, we shall analyze the multiple product control problem discussed above by a simple numerical example.

Let us suppose that the rates of inventory holding cost $h_1(w)$ and $h_2(w)$ are specifically given as follows:

$$h_1(w) = 0.20\,|w| \tag{14}$$

$$h_2(v) = 0.10\,|v| \tag{15}$$

The interest rate is 0.01. Both the rate of inventory holding cost and the interest rate are expressed as monthly rates. We shall consider $T = 20$ months for the planning period, and assume the terminal prices of wheat and rice to be \$15 and \$20, respectively. In this analysis, we shall also assume (for simplicity) that the signs of w and v are always positive, i.e., the firm never runs into a short position for the inventories. If the problem involves changes in the sign of w or v, the problem can be solved by the techniques described in our previous paper [1].

Using the result obtained in the previous paper, we can write the switching function $\mathcal{A}_1$ and $\mathcal{A}_2$ as

$$\mathcal{A}_1 = e^{R(T-t)}(u_1 - p_1) \tag{16}$$

and

$$\mathcal{A}_2 = e^{R(T-t)}(u_2 - p_2), \tag{17}$$

where

$$u_1 \equiv \left[\frac{h_1'(w)}{R} + P_1\right] e^{-R(T-t)} - \frac{h_1'(w)}{R} \tag{18}$$

and

$$u_2 \equiv \left[\frac{h_2'(v)}{R} + P_2\right] e^{-R(T-t)} - \frac{h_2'(v)}{R}. \tag{19}$$

Substituting $T=20$, $h_1'(w)=0.20$, $h_2'(v)=0.10$, $P_1=15$, $P_2=20$, and $R=0.01$, we have

$$u_1 = 35e^{-0.01(20-t)} - 20 \tag{20}$$

and

$$u_2 = 30e^{-0.01(20-t)} - 10. \tag{21}$$

In order to simplify the subsequent computations, let us assume the following values of p_1 and p_2.

$$\begin{cases} p_1 = 35e^{-0.01(20-t)} - t - 20 & \text{for } 0 \leq t \leq 4 \\ p_1 = 35e^{-0.01(20-t)} + t - 28 & \text{for } 4 \leq t \leq 14 \\ p_1 = 35e^{-0.01(20-t)} - t & \text{for } 14 \leq t \leq 20 \end{cases} \tag{22}$$

$$\begin{cases} p_2 = 30e^{-0.01(20-t)} + t - 16 & \text{for } 0 \leq t \leq 12 \\ p_2 = 30e^{-0.01(20-t)} - t + 8 & \text{for } 12 \leq t \leq 19 \\ p_2 = 30e^{-0.01(20-t)} + t - 30 & \text{for } 19 \leq t \leq 20. \end{cases} \tag{23}$$

The numerical values at the end of each month based on (22) and (23) are given for p_1 and p_2 in Table 1.

TABLE 1

End of month	p_1	p_2	End of month	p_1	p_2	End of month	p_1	p_2
0	8.66	8.56	7	9.73	17.34	14	18.96	22.25
1	7.95	9.81	8	11.04	18.60	15	18.29	21.53
2	7.24	11.06	9	12.35	19.87	16	17.63	20.82
3	6.53	12.31	10	13.67	21.14	17	16.97	20.11
4	5.82	13.56	11	14.99	22.41	18	16.31	19.40
5	7.12	14.82	12	16.31	23.69	19	15.65	18.70
6	8.42	16.08	13	17.63	22.97	20	15.00	20.00

The figures in Table 1 are only samples, since p_1 and p_2 must be specified for all values of t and not limited to its integer values in the continuous wheat and rice trading model. It is important here to

emphasize that the above functions for prices are selected only for the sake of simple illustrations and that the solution algorithm described in this paper can be applied to arbitrary price functions.

From (22), (23) and (20), (21), we obtain the values of $u_1 - p_1$, and $u_2 - p_2$ as follows:

$$(24)\qquad \begin{cases} u_1 - p_1 = t & \text{for } 0 \leq t \leq 4 \\ u_1 - p_1 = 8 - t & \text{for } 4 \leq t \leq 14 \\ u_1 - p_1 = t - 20 & \text{for } 14 \leq t \leq 20 \end{cases}$$

$$(25)\qquad \begin{cases} u_2 - p_2 = -t + 6 & \text{for } 0 \leq t \leq 12 \\ u_2 - p_2 = t - 18 & \text{for } 12 \leq t \leq 19 \\ u_2 - p_2 = -t + 20 & \text{for } 19 \leq t \leq 20. \end{cases}$$

We now consider the constraints on the control variables. Suppose that the purchases constraints and the sales constraints are set up as follows:

$$(26)\qquad \begin{cases} m_{1p} + m_{2p} \leq 6. \\ m_{1p} + 2m_{2p} \leq 10 \\ 3m_{1p} + m_{2p} \leq 12 \\ m_{1p}, \quad m_{2p} \geq 0 \end{cases}$$

$$(27)\qquad \begin{cases} m_{1s} + m_{2s} \leq 5 \\ 3m_{1s} + 2m_{2s} \leq 12 \\ m_{1s}, \quad m_{2s} \geq 0. \end{cases}$$

These constraints are shown in Figure 1(a) and 1(b). Note that since (26) and (27) are the only constraints, this problem is a separable control problem.

The solutions to the linear programming problems discussed earlier must take on one of the extreme points shown in the above figure or on a point of a convex combination of two adjacent extreme points. Which extreme point is selected depends upon the ratio of $\frac{\lambda_1}{\lambda_2}$. For example, if $\frac{\lambda_2}{\lambda_1} = 1$, then (2, 4) and (3, 3) are equally preferred since $2\lambda_1 + 4\lambda_2 = 3\lambda_1 + 3\lambda_2$. But if $\frac{\lambda_2}{\lambda_1} > 1$ then (2, 4) becomes better. As $\frac{\lambda_2}{\lambda_1}$ becomes larger and larger eventually (0, 5) is preferred to (2, 4) which occurs when $\frac{\lambda_2}{\lambda_1} > 2$. Similarly as $\frac{\lambda_2}{\lambda_1} < 1$, (3, 3) is preferred to (2, 4) and as $\frac{\lambda_2}{\lambda_1} < \frac{1}{3}$, (4, 0) becomes preferable to (3, 3). If λ_1 and λ_2 are both negative, the sales constraints become effective. If $\frac{\lambda_2}{\lambda_1} > 1$, (0, 5) is better than (2, 3), but as $\frac{\lambda_2}{\lambda_1} < 1$ the latter is preferred until $\frac{\lambda_2}{\lambda_1} < \frac{2}{3}$ at which time (4, 0) is preferred to (2, 3).

These may be summarized in Figure 2. The numbers in brackets show the optimum values of the control variables, the first number for m_{1p} or m_{1s} and the second number for m_{2p} or m_{2s}. If the number is positive, then that number is the value of m_{1p} or m_{2p}, depending upon the element, and m_{1s} or m_{2s} is set equal to zero. If the number is negative, the negative of that number is the value of m_{1s} or m_{2s}, depending upon the element, and m_{1p} or m_{2p} is set equal to zero.

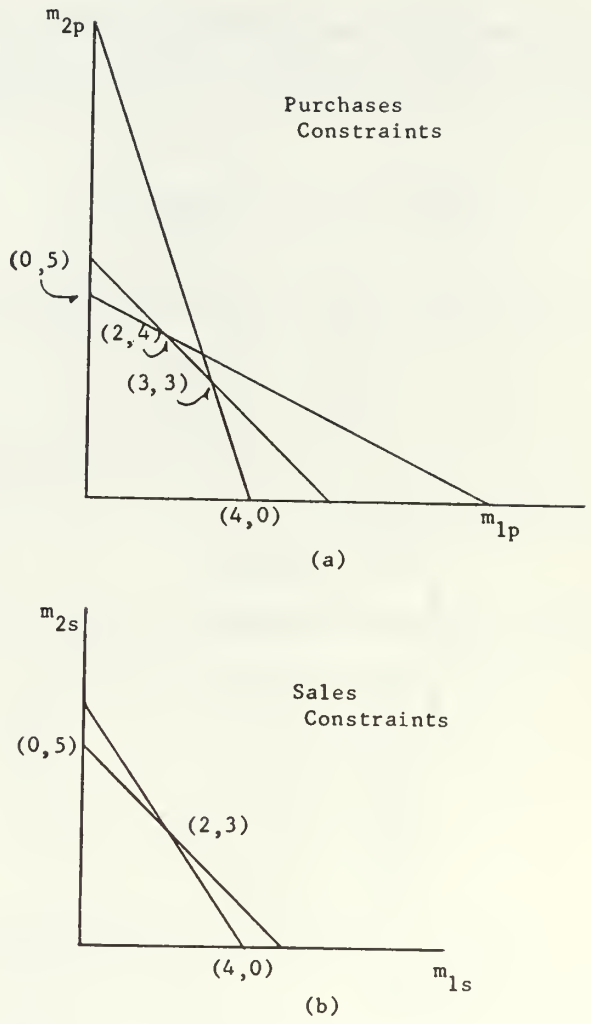


FIGURE 1

Now we relate this analysis with the values of $u_i - p_i$ $(i=1, 2)$ which obtained earlier. Since only the ratio of λ_1 and λ_2 is crucial, and since $\frac{\lambda_2}{\lambda_1}=\frac{u_2-p_2}{u_1-p_1}$, we may use u_1-p_1 and u_2-p_2 in place of λ_1 and λ_2. Then with the equations in (24) and (25) together with Figure 2, we obtain Figure 3. The broken line indicates the change in u_1-p_1 and u_2-p_2 as t changes. The numbers represent the month.

This figure completely specifies the optimum strategy. For example, during Month 2, $\frac{u_2-p_2}{u_1-p_1} > 2$, hence we should purchase only rice at the rate of 5 units per month. However, $\frac{u_2-p_2}{u_1-p_1}$ decreases over time and becomes equal to 1 at the end of Month 2, hence the strategy for Month 3 is to purchase 2 units of wheat and 4 units of rice. Similarly, in Month 13 we should sell only rice at the rate of 5 units a month and Month 14 we should sell 2 units of wheat and 3 units of rice per month.

During the last 2 months sales of wheat at the rate of 4 units a month and purchases of rice at the rate of 5 units a month are required. These are summarized in Table 2.

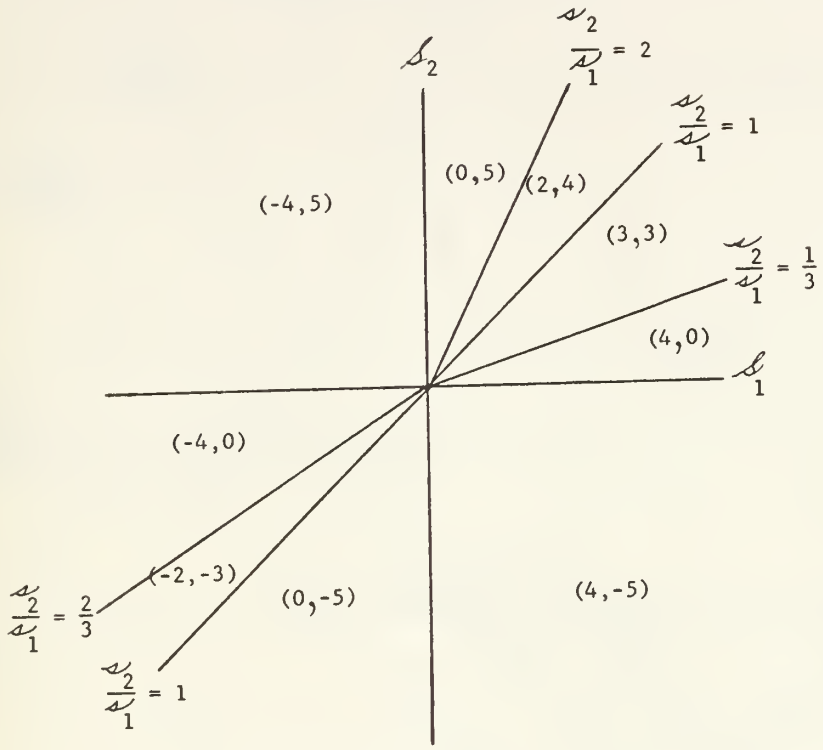


FIGURE 2

TABLE 2

Optimum monthly purchase or sales rate		
Months	Wheat	Rice
1, 2	No trade	Buy 5 units
3	Buy 2 units	Buy 4 units
4, 5	Buy 3 units	Buy 3 units
6	Buy 4 units	No trade
7, 8	Buy 4 units	Sell 5 units
9–13	No trade	Sell 5 units
14	Sell 2 units	Sell 3 units
15–18	Sell 4 units	No trade
19, 20	Sell 4 units	Buy 5 units

In the above example, we designed the prices, for the sake of simple illustrations, in such a way that changes in the control variables occur only at the end of each month. Generally, this is not true, and mid-period corrections in the control variables are often required in continuous control problems.

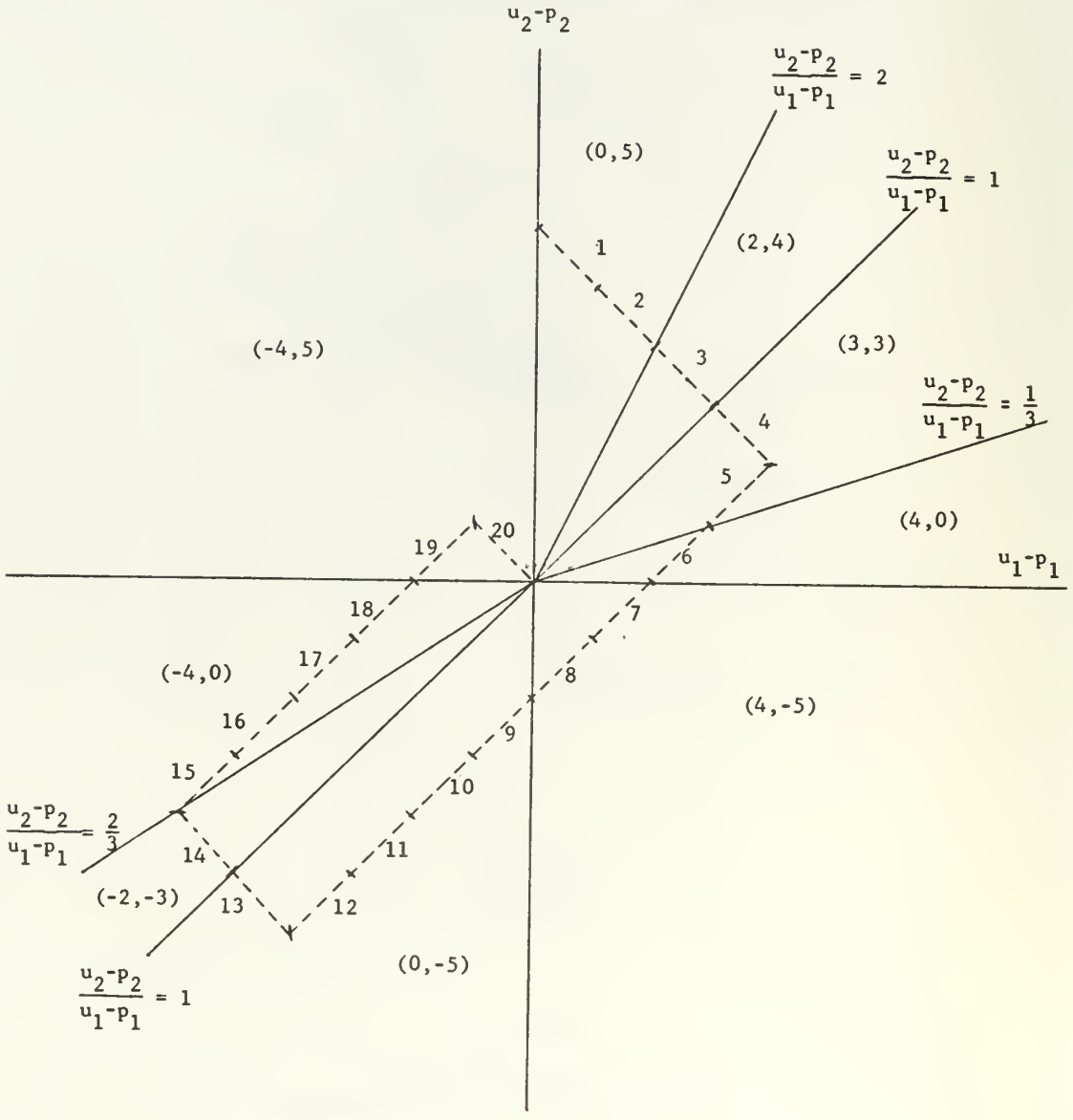


FIGURE 3

4. INTERPRETATIONS OF THE OPTIMAL POLICY

Since the derivations and interpretations of the formulas for the adjoint functions ϕ and ρ, the switching functions $\mathscr{S}_1$ and $\mathscr{S}_2$, and the Hamiltonian function $\mathscr{H}$ are virtually identical to those given in our previous paper [1], we shall only briefly discuss them here.

To obtain the differential equation for Ψ we calculate

$$\dot{\Psi} = -\frac{\partial \mathscr{H}}{\partial c} = -(\Psi - 1)R.$$

The solution to this equation with boundary condition $\Psi(T) = 0$ is

(28) $$\Psi(t) = 1 - e^{R(T-t)}.$$

The interpretation of $\Psi(t)$ is that it is *the net increase in value of a dollar invested from time* t *to* T *at interest rate* R.

The differential equations for ϕ and ρ are obtained as follows:

(w) $$\dot{\phi} = -\frac{\partial \mathscr{H}}{\partial w} = (\Psi - 1)h_1'(w)$$

$$\dot{\rho} = -\frac{\partial \mathscr{H}}{\partial v} = (\Psi - 1)h_2'(v).$$

Substituting (28) into these equations and integrating, we obtain

(29) $$\phi = \int_t^T h_1'(w)e^{R(T-\tau)}d\tau$$

and

(30) $$\rho = \int_t^T h_2'(v)e^{R(T-\tau)}d\tau.$$

As in [1] the interpretation of (29) or (30) is *the total time-adjusted unit cost of storing a small increment of wheat or rice from time* t *to* T.

Substituting these results into the switching functions (10) and (11) we have

(31) $$\sigma_1 = P_1 - p_1e^{R(T-t)} - \phi$$

and

(32) $$\sigma_2 = P_2 - p_2e^{R(T-t)} - \rho.$$

The interpretation of σ_1 is *the future* (*time* T) *price of wheat*, P_1, *less the future value of* $p_1e^{R(T-t)}$ *of paying now* (*time* t) *for wheat at price* p_1, *less the total storage cost* ϕ *from time* t *to* T *of the additional wheat bought now*. The interpretation of σ_2 is similar.

Substituting the results into (9) gives the following form for the Hamiltonian:

(33) $$-\mathscr{H} = [Rc - h_1(w) - h_2(v)] - \sigma(m_p - m_s).$$

From this we can interpret the negative of the Hamiltonian as follows: *the function* $-\mathscr{H}$ *is the future* (*time* T) *value of the present* (*time* t) *cash flow and consists of two parts: the first part is the cash flow due to previous* (*before time* t) *decisions, and the second part is the cash flow due to the present* (*time* t) *decisions.*

It is obvious that we can extend these results to much more general problems with many more control variables, and more complicated constraints. We shall carry out such a generalization in a later paper in this series.

One final interpretation is now possible concerning the dual variables to the maximizing linear programming problem (13). It is well known that the dual to a maximizing problem is a minimizing problem and that one of these problems has a finite solution if and only if the other does. The interpretation of the dual solution depends upon the specific units of the various quantities involved. As an example, assume that A_p, A_s, D_p, and D_s in (13) are pure numbers. Since the units of the switching functions $\mathcal{S}_1$ and $\mathcal{S}_2$ and measured in dollar per unit and m_p and m_s are in units per time period we conclude that b_p, b_s, and b_d have units per time period also. Hence the dual variables to the linear programming problem (13), call them $z_1, \ldots, z_m$, have units of dollar per time period. This means that $z_1(t)$ for example, measures the *rate of contribution to the final asset position if we relax the righthand side of the first constraint by one unit now* (*time* t). This is a rather remarkable interpretation for it evaluates an action now in terms of future values.

REFERENCES

[1] Ijiri, Yuji, and Gerald L. Thompson, "Applications of Mathematical Control Theory to Accounting and Budgeting (The Continuous Wheat Trading Model)," The Accounting Review 45, 246–258 (Apr. 1970).

[2] Pontryagin, L. S., R. V. Boltyanski, R. V. Gamkraldze, and E. F. Mishchenko, *The Mathematical Theory of Optimal Processes* (John Wiley and Sons, New York, 1962).

TOWARD OPTIMAL BIDDING STRATEGIES

Andrew C. Stedry and H. David Brecht

The University of Texas at Austin

ABSTRACT

In this paper we investigate a form of rational behavior in response to an oligopoly pricing problem where only one buyer is involved. We investigate the problem from the standpoint of the seller who wants to maximize his gain from the transaction. In particular, we deal with the problem of one seller's response to an invitation to submit a sealed (i.e., noncooperative) bid to a government or other dominant purchasing agency for supplying a specified bundle of goods and services for which either (1) no other demand exists, or (2) the terms or quantities involved cannot, at least in the short run, be obtained from another source.

Although treated from a normative standpoint, i.e., what bid the supplier should make, the paper also has implications for the buyer's behavior and oligopoly-monopsony* pricing in a more general sense.

INTRODUCTION

The bargaining situation which exists in most government contracting is essentially that of a monopsonist bargaining with an oligopoly. The government is the sole customer for a segment of the goods or services of a (generally small) number of firms. Departures from this norm exist, but government contractors usually have some portion of their facilities which cannot be conveniently converted for nongovernmental use. Such departures serve only to complicate the picture in any event; the situation becomes that of an oligopoly dealing in two markets, one of which is a semi-monopsony.

The process is repeated at lower levels. Prime contractors must select from among a few subcontractors. In addition, the problem exists for contracting with certain nongovernment organizations. Automobile manufacturers operate with many of their suppliers in markets resembling monopsony-oligopoly confrontations; and the suppliers of "house brand" merchandise for nationwide retailing organizations are similarly affected. Bidding for concessions—e.g., food and full services on limited access highways, food and beverage franchises at recreational facilities—are examples which involve a nonfederal government of semi-monopsony.

Despite developments in game theory evolving from von Neumann and Morgenstern's [10] initial formulation to Harsanyi's [3] recent treatise on n-person, incomplete information, etc., games, the application on the theory at the two-person level to a real monopoly-monopsony bargaining problem remains largely a matter of "relative power" of both sides.† Also, the bargaining power of a few dealing with many has been analyzed at length by Fellner [1]. The problem of one dealing with a few has been relatively untouched.

---

*A monopsony market is one in which there is a single buyer who determines price and multiple sellers competing to satisfy the demand.

†See e.g., J. R. Hicks [4] on union-management bargaining, the classic monopoly-monopsony case. His conclusions are in essence that each side will make a concession when it costs less than a strike would.

Nash's [6] important work provides some insights, but in the contracting problem the dominan player and the oligopoly members are really playing different games. Team theory, as developed b Marschak [5] and Radner [7], provides insights. However, the coalitions found in both n-persons game and team decisions are illegal in most government contracting competition and considered bad form a best in private endeavors. Vickery's work [9] with competitive bidding is of interest, but his method for achieving Pareto-optimality necessitate significant perturbations of current bidding procedures

WHY NOT DISCRIMINATING MONOPSONY?

In either the governmental or the private situation it might be supposed that the discriminatin monopsonist could simply select the most efficient producer for each bundle of goods or services i desired. Two factors serve as obstacles to this approach. First, the monopsonist does not know, priori, who the most efficient producer is at any given time. Short run slack is a major factor affect ing a firm's capability and willingness to take a large order. Second, some mechanisms resembling th mirror image of the behavior of a discriminating monopolist in a market with numerous buyers—i.e. selecting an arbitrarily low price and moving upward until the first "taker" is found—would undoubt edly create a monopoly in one form or another. If the process is successful, ultimately all but th most efficient producer would be driven from the market, thus leaving the classic monopoly-monopsony case.

More likely, however, is the creation of a monopoly in the form of a coalition, the prevention of which rests on legal sanctions strongly upheld by concepts of fairness.

Furthermore, in practice, the discriminating monopsony attempt would take an indeterminate amount of time for coalitions to form, for slack to disappear, and in general, for the monopsonist to be pressed for a decision as his requirement for the goods or services becomes critical.

The compromise solution reached in practice is the sealed competitive bid. The process restricts shared information flow; a discriminating monopsonist would divulge considerable information with his first offer. There is no particular reason for believing that the lowest of the sealed bids would be higher than the price at which the same bidder would accept the published monopsony price. Indeed, the successful bidder wins on his offer, not that of the buyer.

Finally, the sealed bid appears to satisfy notions of equity to the extent that gross violations of the noncoalition provisions are relatively infrequent. It is used, in any event, by all levels of government and numerous nongovernment agencies in the belief that it represents the best means of achieving minimum longrun cost.

We depart here from the work of Rothkopf [8] and Griesmer and Shubik [2] in that we assume that benefit can be gained from use of an optimal technique where these authors do not. Moreover, as will become apparent, we make fewer assumptions about necessary bidding distributions.

MODEL BIDDING ENVIRONMENT

We assume that each bidding opportunity is an independent entity.* The contract basis is fixed price, i.e., the successful bidder receives the amount he bids from the buyer, supplying the requisite goods or services without recourse for cost overruns, etc. Associated with satisfying the con-

*Since we are dealing with small samples, the work done by J. E. Walsh with order statistics' independence in large samples is not relevant.

ract is a cost incurred if the bid is won and not otherwise. No long-run fixed costs are considered; ve assume simply that if accepting a bid will necessitate increased capacity, the net cost of the additional investment after a suitable deduction for post-contract value will be added to the contract ost.

We assume that all bids are comparable and that the buyer will accept the lowest. First, consider he case where competitors' bids are considered to be a sample drawn from an arbitrary distribution.

At this point we assume only that the distribution exists and is continuous. We consider the bidding nvironment faced by one bidder who has a known number, n, of competitors.

Let x_j be the bid price submitted by the jth bidder. The sample then consists of $x_1, \ldots, x_n$, the prices bid by the n competitors. Without loss of generality the bids can be reordered so that $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$. These then are the order statistics of a sample of size n. It is well known* that the cumulants of the order statistics, $F(x_{(j)})$ are themselves random variables which are beta-distributed with parameters j and $n-j+1$. Following Wilks, we designate this by $Be\ (j, n-j+1)$. Recalling that a random variable, say y, which is $Be\ (v_1, v_2)$ has density given by

$$f(y) = \frac{\Gamma(v_1+v_2)}{\Gamma(v_1)\Gamma(v_2)} \quad y^{v_1-1} \quad (1-y)^{v_2-1}, \tag{1}$$

and that $\Gamma(\alpha+1) = \alpha!$ for integral values of α we obtain for the density of $F(x_{(j)})$

$$\psi[F(x_{(j)})] = \frac{n!}{(j-1)!\,(n-j)!} \quad [F(x_{(j)})]^{j-1}\,[1-F(x_{(j)})]^{n-j}. \tag{2}$$

The cumulants may be considered in terms of the *coverages* of the parent distribution represented by the order statistics, i.e., the actual amounts of probability lying between adjacent points represented in the sample and on either side of the sample. The coverages $u_1, \ldots, u_{n+1}$ are defined by $u_1 = F(x_{(1)})$, $u_2 = F(x_{(2)}) - F(x_{(1)}, \ldots, u_{n+1} = 1 - F(x_{(n)})$. The sum of *any* k of the coverages is known to be $Be\ (k, n-k+1)$, a fact that could be relevant in extensions where the contract is shared by two or three of the lowest bidders. It is readily verified that (2) is a special case where the first j are summed to yield $F(x_{(j)})$. A more relevant special case here is that the density of any of the coverages is given by:

$$\varphi(u_j) = n(1-u_j)^{n-1}. \tag{3}$$

The $(n+1)$st bidder will be successful if he selects a bid $x_{n+1} < x_{(1)}$. If $F(x)$ is also assumed to be monotone increasing in x, $x_{n+1} < x_{(1)} \langle = \rangle F(x_{n+1}) < F(x_{(1)})$. Hence, for any x_{n+1} selected there corresponds a $\tilde{u} = F(x_{n+1})$ such that the probability of a successful bid can be expressed in terms of the distribution of u_1:

$$P\{x_{n+1} < x_{(1)}\} = P\{\tilde{u} < u_1\} = 1 - P\{u_1 < \tilde{u}\}. \tag{4}$$

But, from (3)

*See, e.g., Wilks [11], pp. 173, 234–245.

$$P\{u_1 \leq \tilde{u}\} = \Phi(\tilde{u}) = \int_0^{\tilde{u}} n(1-u_1)^{n-1} du_1 = 1-(1-\tilde{u})^n. \tag{5}$$

Combining (4) and (5), we obtain

$$P\{x_{n+1} < x_{(1)}\} = (1-\tilde{u})^n. \tag{6}$$

In order to obtain a desired probability of success, say $\tilde{P}$, (6) may be solved directly for $\tilde{u}$.

$$\tilde{u} = 1 - n\sqrt{\tilde{P}}. \tag{7}$$

As a result of the monotonicity assumption a unique solution of $\tilde{u} = F(x_{n+1})$ exists. Finding it, o course, requires that the distribution function be known, at least in the relevant range. We shall now consider the effects of various assumptions involving the probabilities of success and the nature o the bidding decision.

NO INFORMATION

Suppose the $(n+1)$st bidder has no information about the decision. Then, presumably, he is simply drawing another independent sample, of size 1, from the parent population. Using primes and a subscript to denote the sample of 1,

$$\varphi 1'(u_1') = 1 \cdot (1-u_1')^{1-1} = 1. \tag{8}$$

Since the samples are independent,

$$\varphi(u_1, u_1') = \varphi(u_1)\varphi_1(u_1') = n(1-u_1)^{n-1}. \tag{9}$$

The probability of a successful bid is given by (4) with u_1' substituted for $\tilde{u}$. Here, however, u_1' is a random variable also:

$$P\{u_1' < u_1\} = \int_0^1 \int_0^{u_1} n(1-u_1)^{n-1} du_1' du_1 = \frac{1}{n+1}. \tag{10}$$

The same result could have been obtained by reasoning that in the absence of information the $(n+1)$st bid is one member of a sample of size $n+1$, each member having an equal likelihood of being $x_{(1)}$. However, the method used above can be generalized.

Suppose m independent groups within the $(n+1)$st firm arrive at m independent estimates for x_{n+1} and the lowest of these is chosen. The relevant coverage is now u_1' for a sample of size m, and by the use of subscripts to denote the interval sample size, (8) becomes:

$$\varphi_m'(u_1') = m(1-u_1')^{m-1}. \tag{11}$$

The generalizations of (9) and (10) are then

$$\varphi(u_1, u_1') = (u_1)\varphi_m(u_1') = n(1-u_1)^{n-1}m(1-u_1')^{m-1} \tag{12}$$

and

$$P\{u_1' < u_1\} = \int_0^1 \int_0^1 n(1-u_1)^{n-1}m(1-u_1')^{m-1}du_1'du_1 = \frac{m}{m+n}. \tag{13}$$

Thus, assuming that the m groups are truly independent and operating under the same criteria as would a random sample of the competitors, the probability of a successful bid could be made arbitrarily large (as long as the competitors were not doing the same thing).

Another interesting extension, and a more typical case, is bidding the median of m intrafirm independent groups. Consider $m=3$. The relevant cumulant is $F(x_{(2)})$ for a sample of size 3. Recalling (2) and letting $u''=F(x_{(2)})$, we have

$$\psi_3(u'') = (3!)(u'')(1-u''), \tag{14}$$

$$\Psi_3(u_1) = \int_0^{u_1} (3!)(u'')(1-u'')du'' = 3u_1^2 - 2u_1^3, \text{ and} \tag{15}$$

$$P\left\{u'' < u_1\right\} = \int_0^1 \int_0^{u_1} (3!)(u'')(1-u'')n(1-u)^{n-1}du''du_1 = \frac{6}{(n+2)(n+3)}. \tag{16}$$

Note that this probability is *less* than $\frac{1}{n+1}$ for $n > 1$. The process of taking the median estimate, rather than a single estimate, within a firm reduces the probability of a successful bid. On reflection this is not surprising. The intrafirm median is an estimate of the median of the parent and, indeed, it is easily shown that $E(u'')=1/2$. If the firm knows the median of the parent population and bids the median, its probability of success is given by (6) with $\tilde{u}=1/2$ or $(1/2)^n$, whose value is less than $\frac{1}{n+1}$ for $n > 1$, increasingly so as n increases. (An analogy is where the random bid is equivalent to drawing straws with n other players. By contrast, bidding the median gives each player an even chance of being lower. It is akin to flipping a coin with each player in turn and being successful only if he wins every toss.)

Obviously, this line of analysis could be extended to other intrafirm arrangements as well as order statistics of other samples—e.g., prior bids—which are assumed to come from the same parent distribution. Instead, however, we shall investigate the effects of differing assumptions about the parent population on a rational bidding procedure.

EXPECTED PROFIT

Earlier we discussed the cost structure considered. The cost, say c_j, is the cost to the jth bidder if he is successful, independent of the bid price. Clearly the bid price is his revenue, if successful; both are zero if the bidder is not successful.

We introduce a new random variable, w_j, whose value is 1 if the bid is successful and 0 otherwise. Then the jth firm's profit can be expressed as

$$\pi_j = (x_j - c_j)w_j, \qquad j = 1, \ldots, n+1. \tag{17}$$

If expected profit is desired, we can permit c_j to be a random variable without complication provided its mean, say $\bar{c}_j$, is known and c_j and w_j are stochastically independent. Observe that, to the jth bidder, x_j is a decision variable and not a random variable. We can then express expected profit as

$$\bar{\pi}_j = E(\Pi_j) = (x_j - \bar{c}_j)E(w_j). \tag{18}$$

But

$$E(w_j) = 0 \cdot P\{w_j = 0\} + 1 \cdot P\{w_j = 1\} = P\{w_j = 1\}. \tag{19}$$

Rewriting (18), we obtain

$$\bar{\Pi}_j = (x_j - \bar{c}_j)P\{w_j = 1\}. \tag{20}$$

The probability that $w_j = 1$ is, of course, the probability of a successful bid. It is a function of the coverage of the bidding distribution by the bid price x_j and hence of x_j if the parent distribution of bids is known. Therefore, designate this function as $P(x_j)$.

If we now assume that the range of possible bids is $a \leq x_j \leq b$ and that $P(x_j)$ is twice differentiable in $a < x_j < b$, $\bar{\Pi}_j = (x_j - \bar{c}_j)P(x_j)$ is at a maximum where a unique solution, x_j exists to

$$\frac{d\bar{\Pi}_j}{dx_j} = (x_j - \bar{c}_j)P'(x_j) + P(x_j) = 0, \tag{21}$$

and that

$$\left.\frac{d^2\bar{\Pi}_j}{dx_j^2}\right|_{x_j^*} = (x_j^* - \bar{c}_j)P''(x_j^*) + 2P'(x_j^*) < 0. \tag{22}$$

RECTANGULAR BID DISTRIBUTION

We now assume that the $(n+1)$st bidder knows that his competitors' bids will be randomly distributed so that any bid over a range, r, such that all x_j, $d \leq x_j \leq d + r$ are equally likely. For convenience we designate $x_{n+1} = x$, $\bar{c}_{n+1} = \bar{c}$ for the following discussion.

The coverage of this parent distribution represented by a bid x is easily derived from

$$f(x) = \begin{cases} 1/r & d \leq x \leq d + r \\ 0 & \text{elsewhere} \end{cases} \tag{23}$$

and

$$F(x) = \int_d^x \frac{1}{r}\, dj = \frac{x - d}{r} \quad d \leq x \leq d + r.$$

From (6) and (23) the probability of x being a successful bid is

$$P(x)=\begin{cases}0 & x>d+r\\ \left(1-\dfrac{x-d}{r}\right)^n & d\leq x\leq d+r.\\ 1 & x<d\end{cases} \tag{24}$$

We note that $P(x)$ is twice differentiable in the closed interval $d\leq x\leq d+r$ and that no gain in expected profit can be obtained by a bid $x<d$ rather than $x=d$ or $x>d+r$. Therefore, we can restrict one attention to bids within the range.

Substituting in (21) we obtain

$$\frac{d\bar{\Pi}}{dx}=-(x-\bar{c})\,\frac{n}{r}\left[1-\left(\frac{x-d}{r}\right)\right]^{n-1}+\left[1-\left(\frac{x-d}{r}\right)\right]^{n}. \tag{25}$$

Taking the second derivative, for $n>1$ we obtain

$$\frac{d^2\bar{\Pi}}{dx^2}=(x-\bar{c})\,\frac{n(n-1)}{r^2}\left[1-\left(\frac{x-d}{r}\right)\right]^{n-2}-\frac{2n}{r}\left[1-\left(\frac{x-d}{r}\right)\right]^{n-1}. \tag{26}$$

Two distinct solutions exist to $d\bar{\Pi}/dx=0$, namely,

$$\left[1-\frac{x-d}{r}\right]^{n-1}=0 \tag{27a}$$

and

$$\frac{-n(x-\bar{c})}{r}+\left[1-\frac{x-d}{r}\right]=0. \tag{27b}$$

Solving for x in (27a) yields $x=d+r$, a bid that yields $P(x)=0$ and $E(\Pi)=0$. The second derivative evaluated at $x=d+r$ is, of course, 0. This point is therefore not a true maximum. It is an optimal solution only if any $x<d+r$ produces negative profit, i.e., only if $\bar{c}\geq d+r$.

Solving (27b) for x yields

$$x=\frac{n\bar{c}+d+r}{n+1}=\bar{c}+\frac{(d-\bar{c})+r}{n+1}. \tag{28}$$

It is obvious from the second form for x in (28) that $\bar{c}>d+r$ would yield $x<c$, the situation in which (27a) provides a better solution. On the other end of the bidding range a bid $x<d$ will be obtained only if

$$\frac{n\bar{c}+d+r}{n+1}<d,$$

or

$$\bar{c}<d-\frac{r}{n}. \tag{29}$$

Evaluating the second derivative where x is given by (28) yields

$$\frac{d^2\Pi}{dx^2} = -\frac{n}{r}\left[\frac{n(d+r-\bar{c})}{r(n+1)}\right]\left[\frac{d+r-\bar{c}}{r}\right], \tag{30}$$

which is clearly negative for $d+r>c$.

We can therefore summarize the optimal bidding procedure as

$$x^* = \begin{cases} d & \text{where } \bar{c} < d-\dfrac{r}{n}, \\ \bar{c}+\dfrac{(d-\bar{c})+r}{n+1} & \text{where } d-\dfrac{r}{n} \leq \bar{c} \leq d+r, \text{ and} \\ d+r & \text{where } \bar{c} > d+r. \end{cases} \tag{31}$$

The expected profit is readily found to be

$$\bar{\Pi}(x^*) = \begin{cases} d-\bar{c}\bar{c} < d-\dfrac{r}{n} \\ \dfrac{r}{n}\left[\dfrac{n(d+r-\bar{c})}{r(n+1)}\right]^{n+1} & d-\dfrac{r}{n} \leq \bar{c} \leq d+r \\ 0 & \bar{c} > d+r. \end{cases} \tag{32}$$

Some insight can be gained by considering the case where $d=\bar{c}$. This insight arises if the $(n+1)$st firm is among the most efficient firms so that no bids would appear below its average cost. Then the solutions would be

$$x^* = \bar{c}+\frac{r}{n+1},$$

$$P(x^*) = \left[\frac{n}{n+1}\right]^n, \tag{33}$$

and

$$\bar{\Pi}(x^*) = \frac{r}{n+1}\left[\frac{n}{n+1}\right]^n.$$

Here, the $(n+1)$st firm would bid so that the bid's coverage of the bidding distribution is exactly $1/(n+1)$. Its available profit, if the bid is successful, is proportional to the bidding range, r.

ARBITRARY KNOWN DISTRIBUTIONS

We now consider a class of bidding distributions for which $\bar{u}(x)=F(x)$ in a known continuous function in an interval $a \leq x \leq b$ with known and continuous derivatives of order two in $a<x<b$ such that $F(y)$, $y=x-\bar{c}$, has the same properties in $a-\bar{c} \leq y \leq b-\bar{c}$. Such a linear transformation will usually not produce undue difficulty. In fact, if the $(n+1)$st bidder is among the most efficient producers, a reasonable assumption in many instances is that the range of the bidding distribution is

bounded below by $\bar{c}$, thus enabling $F(y)$ to be more simply represented than $f(x)$ by one of the common densities—e.g., gamma, log normal, exponential—which are typically defined for $0 \leq y \leq \infty$.

Recalling (6), the probability of a successful bid is then given by

$$P\{y < y_{(1)}\} = [1-F(y)]^n. \tag{32}$$

Expected profit and its derivatives can then be derived in terms of $F(y)$, $f(y) = F'(y)$ and $f'(y)$ as

$$\bar{\Pi} = y[1-F(y)]^n$$

$$\frac{d\bar{\Pi}}{dy} = -ny[1-F(y)]^{n-1}f(y) + [1-F(y)]^n \tag{34}$$

$$\frac{d^2\bar{\Pi}}{dy^2} = -n\{yf'(y)[1-F(y)]^{n-1} + 2f(y)[1-F(y)]^{n-1} - (n-1)y[f(y)]^2[1-F(y)]^{n-2}\}.$$

Analogous to the rectangular distribution case, any bid $y < b-\bar{c}$ will yield $F(y) < 1$ so that if it is worth bidding at all the optimality conditions are found, after some manipulation, to be

$$nyf(y) = [1-F(y)] \tag{35a}$$

$$nyf'(y) + (n+1)f(y) > 0. \tag{35b}$$

We now see that the rectangular distribution is a special case of any coverage which can be represented approximately by

$$F(y) = A + By, \tag{36a}$$

in a relevant range $C \leq y \leq D$ where "relevant" is defined by having a solution for y in that range such that (since $f(y) = B$)

$$nBy = 1 - (A + By)$$

Yielding the solution for y, viz y^*, as:

$$y^* = \frac{1-A}{B(n+1)} \tag{36b}$$

Clearly, for $A = 0$ and $B = 1/r$, (36b) yields (33) with $y^* = x^* - \bar{c}$. Also, for $F(y)$ linear $f(y) = 0$ so the second derivative condition (35b) is fulfilled.

For contrast, we consider a highly skewed distribution where y is exponentially distributed, say $f(y) = \lambda e^{-\lambda y}$, $0 \leq y \leq \infty$, $F(y) = 1 - e^{-\lambda y}$ and (35a) becomes

$$ny\lambda e^{-\lambda y} = e^{-\lambda y}, \tag{37a}$$

whose solution is

$$y^* = 1/n\lambda, \tag{37b}$$

which yields

$$F(y^*) = 1 - e^{-1/n} \tag{37c}$$

and

$$P(y^*) = [1 - F(y^*)]^n = e^{-1}. \tag{37d}$$

The optimal probability of a successful bid is independent of the number of bidders; however, this probability is the limiting case for the rectangular distribution with $\bar{c} = d$ as $n \to \infty$. Recalling (33), we see that:

$$\lim_{n\to\infty} P(x^*) = \lim_{n\to\infty} \left[\frac{n}{1+n} \right]^n = \lim_{n\to\infty} \frac{1}{\left(1+\frac{1}{n}\right)^n} = e^{-1}. \tag{38}$$

The exponential is a highly (positively) skewed distribution with the probable competitive bids clustered quite close to zero. By contrast, the rectangular distribution is not skewed at all. A bidding distribution which is negatively skewed is difficult to envision in practice. Yet the optimal probability of a successful bid for the rectangular distribution (over $\bar{c} \leq x \leq \bar{c} + r$) lies in a quite narrow range, from 1/2 when $n = 1$ to $1/c$ as $n \to \infty$. For the exponential distribution, independent of n (and, incidentally, of its mean) optimal probability is $1/c$. This suggests the plausibility of the following: an optimal bid is one which produces a probability of success in the neighborhood of $1/c$ or 37 percent in "highly competitive" situations, growing to about 50 percent where competition is minimal.

For $n \geq 2$ these numbers are higher than the proportions that would be gained under a "fair share" division with each bidder winning the bid with probability $1/n+1$.

EFFECT ON THE BUYER

Again consider a rectangular density of bids $f(x) = 1/r$, $\bar{c} \leq x \leq \bar{c} + r$ which is equivalent to $f(y) = 1/r$, $0 \leq y \leq r$. It can be shown that the sampling distribution of the lowest of n bids, $y_{(1)}$, is given by*

$$\varphi_n(y_{(1)}) = \frac{n}{r}\left(1 - \frac{y_{(1)}}{r}\right)^{n-1}. \tag{39}$$

If all $n+1$ bidders sample at random from the density, buyer's expected price is given by $E(y_{(1)})$ for a sample of $n+1$, viz.

$$E(y_{3|4}) = \int_0^r y_{(1)}\Phi_{n+1}(y_{(1)})\, dy_{(1)} = \frac{r}{n+2}. \tag{40}$$

*Since $u_1 = \frac{y_1}{r}$, performing that transformation in (3) gives the result immediately.

Thus, if no bidder behaves optimally in an expected value sense, the sealed competitive bid will produce a bid which covers a proportion $1/(n+2)$ of the range. Viewing $\bar{c}+r$ as the cost of fulfilling the contract for the least efficient firm, the range, r, could be viewed as the quasi-rent obtainable by the efficient successful bidder in a pure negotiation. The sealed bid competition shares this quasi-rent quite unevenly, with the buyer obtaining $n+1$ parts, the seller only 1, reinforcing the conjectures made in the introductory sections about the efficacy of the sealed bid.

Where one bidder behaves optimally, the price paid by the buyer will be the optimal bid, $y^*=r/(n+1)$ where $y_{(1)}$ from a sample of n bids exceeds y^*, but $y_{(1)}$ otherwise. Let this price be designated by y. Then

$$E(y)=\int_0^{r/n+1} \cdot\phi_n(y)dy+\int_{r/n+1}^{r} r/n+1\phi_n(y)dy \tag{41}$$

$$=\frac{r}{n+1}\left[1-\frac{1}{n+1}\left(\frac{n}{n+1}\right)^n\right].$$

The buyer's average price will be increased, although not appreciably. For $n=2$, (41) yields $0.284r$ rather than $0.25r$; with six bidders, i.e., $n=5$, (41) yields $0.155r$ instead of $0.143r$ given by (40). The buyer's price can never exceed $r/(n+1)$; however, so that his additional cost might be compensated for by the assurance that in no case will price exceed $0.333r$ with three bidders or $0.167r$ with six. At worst, the buyer's price would increase to these limits if all bidders behaved optimally. Also to be noted is the fact that the buyer's price in all cases investigated approaches the cost of the most efficient producer as the number of bidders increases, i.e., approaches a purely competitive market.

GAIN FROM OPTIMAL BIDDING

For the rectangular distribution $\bar{c} \leq x \leq \bar{c}+r$ we found that (33):

$$\overline{\Pi}(x^*)=\frac{r}{n+1}\left[\frac{n}{n+1}\right]^n. \tag{42}$$

From (40) we see that a random bid will yield a lower average profit, viz. $r/(n+2)$ and a lower probability of success, $y_{(n+1)}$. Thus, for a random bid x^r, expected profit is given by

$$\overline{\Pi}(x^r)=\frac{r}{(n-1)(n+2)}. \tag{43}$$

The ratio of these two expressions is

$$\frac{\overline{\Pi}(x^*)}{\overline{\Pi}(x^r)}=(n+2)\left(\frac{n}{n+1}\right)^n. \tag{44}$$

For $n=1$, this ratio is 3/2; $n=2$ provides 16/9 and the ratio increases without bound as n grows large.

SUMMARY AND CONCLUSIONS

We have considered a very simple optimization problem from the standpoint of one bidder in a sealed-bid situation who desires to maximize his expected value on that bid. Using order statistics it was possible to define the criteria for an optimal bid where the first two derivations are continuous. For a rectangular distribution competitive bids, the rational bidder who is also an efficient producer experiences at least a 50 percent gain in expected profit over that obtainable by a randomly selected bid and this gain increases with the number of competitors.

It would appear that from the standpoint of the buyer, a competitive sealed bid, if no coalitions are in fact formed, is a highly effective mechanism. Optimal behavior on the part of one of the bidders increases the buyer's expected price slightly, but probably not significantly. Perhaps what is most important from the buyer's standpoint is that optimal independent behavior, even for all of the bidders, will affect his cost very little; but the formation of coalitions such that the effective number of bidding firms is reduced to one, or even two or three, will raise his costs substantially.

ACKNOWLEDGEMENTS

The authors are indebted to the Corn Products Corporation and the Office of Naval Research, Contract N00014–67–A–0126–0008, who partially supported the work.

REFERENCES

[1] Fellner, W., *Competition Among the Few* (Alfred Knopf, New York, 1949).

[2] Griesmer, J. H. and M. Shubik. "Toward a Study of Bidding Processes," (four parts) Nav. Res. Log. Quart., *10*, 11–21 (1963); *10*, 151–173 (1963); *10*, 199–217 (1963); and *14*, 415–433 (1967).

[3] Harsanyi, J. C., "Games with Incomplete Information Played by 'Bayesian' Players," (three parts) Management Science, *14*, 159–182 (1967); *14*, 320–334 (1968); and *14*, 486–502 (1968).

[4] Hicks, J. R., *The Theory of Wages* (Macmillan Co., London, 1965).

[5] Marschak, J., "Elements for a Theory of Teams," Management Science *1*, 127–137 (1955).

[6] Nash, J. F., "The Bargaining Problem," Econometrica *18*, 155–162 (1950).

[7] Radner, R., "The Application of Linear Programming to Team Decision Problems," Management Science *5*, 143–150 (1959).

[8] Rothkopf, M. H., "A Model of Rational Competitive Bidding," Management Science *15*, 362–373 (1967).

[9] Vickery, W., "Counterspeculation, Auctions, and Competitive Sealed Tenders," J. Finance *16*, 8–37 (1961).

[10] von Neumann, J., and O. Morgenstern, *Theory of Games and Economic Behavior* (Princeton University Press, (Princeton, N.J., 1947), 2d edition.

[11] Wilks, S. S., *Mathematical Statistics* (John Wiley & Sons, New York, 1962).

POSTOPTIMALITY ANALYSIS IN ZERO-ONE PROGRAMMING BY IMPLICIT ENUMERATION

Gary M. Roodman

Amos Tuck School of Business Administration
Dartmouth College

ABSTRACT

The procedures for postoptimality analysis that are so much a part of linear programming studies have no simple counterparts in an integer programming context. In the case of Balas' Additive Algorithm, however, it would appear that the capacity of the technique to facilitate certain types of postoptimality analysis has not been fully exploited. This paper proposes an extension of the additive algorithm that utilizes insights generated while solving the original problem to do subsequent analysis upon it. In particular, procedures are developed for doing limited parameter ranging and for seeking new optima in light of parameter changes.

0. INTRODUCTION

The procedures for postoptimality analysis that are so much a part of linear programming studies have no simple counterparts in an integer programming context. The inherent "lumpiness" of the feasible region (as it has been described by Gomory and Baumol [11]) makes it difficult to do the usual types of ranging analysis, to evaluate shifts in constraints, and to achieve efficient restarting as parameters are changed.* Furthermore, the (in)efficiency of most solution methods in the area renders undesirable, if not infeasible, the more direct approach of solving multiple problems with alternative sets of parameters. In the case of at least one of these methods, however, Balas' Additive Algorithm [2], it would appear that the capacity of the technique to facilitate certain types of postoptimality analysis has not been fully exploited. This paper proposes an extension of the additive algorithm that utilizes insights generated while solving the original problem to do subsequent analysis upon it.

In section I the additive algorithm is briefly reviewed, with emphasis on particular characteristics of the method. A modified additive algorithm is outlined in section II together with procedures for doing limited parameter ranging and for seeking new optima in light of parameter changes. In section III certain extensions and generalizations of the methodology are discussed. The special problems involved in employing composite constraints are examined in section IV. The paper is then concluded with a brief discussion of implementation of the procedures on the time-sharing computer. A numerical example appears in the appendix.

I. THE ADDITIVE ALGORITHM

Consider the pure (0, 1)-problem of the form

*See, however, the work of Gomory and Baumol [11], where dual evaluators are derived which have important properties in common with linear programming dual evaluators. See also Alcaly and Klevorick [1], Balas [4], and Jensen [12].

(I) Minimize $z = \underline{c}\underline{x}$, subject to $A\underline{x} \leq \underline{b}$, $x_j = 0$ or 1,

where $A = ||a_{ij}||$ is an $m \times n$ matrix, $\underline{c} = (c_1, \ldots, c_n)$, $\underline{x} = (x_1, \ldots, x_n)$, and $\underline{b} = (b_1, \ldots, b_m)$. Any $\underline{x}$ such that $x_j = 0$ or 1, $j = 1, \ldots, n$, is called a *solution* to problem (I). A solution is *feasible*, if it satisfies $A\underline{x} \leq \underline{b}$; and if in addition it minimizes z, it is called an *optimal feasible solution*.

A *partial solution*, S, is defined to be an assignment of (0, 1)-values to a subset of the n variables. Variables in the partial solution are termed *fixed*, while all other variables are *free*. When x_j is in S its value (0 or 1) is denoted by x_j^S. We define index sets

$$J^S = \{j | x_j \text{ is fixed with respect to } S\}$$

$$\bar{J}^S = \{j | x_j \text{ is free with respect to } S\}.$$

Any solution formed by assigning (0,1)-values to all x_j, $j \in \bar{J}^S$, is called a *completion* of S.

Problem (I) has 2^n solutions, all of which could in theory be enumerated to find the optimal feasible solution. In general, however, this is not necessary. Balas' additive algorithm is one method for undertaking the enumerative process in a way that permits exclusion of large numbers of solutions without explicit examination.* In the additive algorithm the excluded solutions consist of completions of selected partial solutions. A partial solution is said to have been *fathomed* when it is established that none of its completions require further examination. There are many systematic ways of generating a sequence of partial solutions that effects a nonredundant, but exhaustive search of all solutions (see e.g., Balas [2], Glover [9], Geoffrion [7], and Lemke and Spielberg [14]).

Fathoming is accomplished in the additive algorithm through a series of simple tests. Denote by $\bar{z}$ the value of the best feasible solution so far encountered in the search, and append as constraint 0 $\underline{c}\underline{x} \leq \bar{z}$ ($b_0 = \bar{z}$, $a_{0j} = c_j$, $j = 1, \ldots, n$), thereby rendering infeasible all solutions poorer than the current best. In addition, form the index sets

$$\bar{J}_-^S = \{j | j \in \bar{J}^S \text{ and } c_j < 0\}$$

$$\bar{J}_+^S = \{j | j \in \bar{J}^S \text{ and } c_j \geq 0\}.$$

Then, for the partial solution, S,, under examination, define†

$$t_i^S = b_i - \sum_{j \in J^S} a_{ij} x_j^S, \qquad i = 0, 1, \ldots, m,$$

$$\rho_i^S = \sum_{j \in \bar{J}^S} \min\{a_{ij}, 0\} \qquad i = 0, 1, \ldots, m,$$

*The additive algorithm has been described as a *single-branch* enumerative scheme (see Balinski and Spielberg [5]). The other general class of methods have been characterized as *multi-branch* enumerative schemes. See Land and Doig [13] and Little, Murty, Sweeney, and Karel [15]. In fact, the scheme developed here, as part of the postoptimal phase, falls somewhere between these two classifications.

†For purposes of interpreting these quantities, t_i^S may be viewed as the amount of slack (positive or negative) in constraint i, while $-\rho_i^S$ is the largest increase in t_i^S that can be effected by any completion of S. δ_i^S is the change in t_i^S that would result if S were completed in the most favorable way possible (as measured by the coefficients of the objective function).

$$\delta_i^S = \sum_{j \in \bar{J}_-^S} a_{ij} \qquad i = 0, 1, \ldots, m.$$

We state the following basic tests:*

Test 1: If $\delta_i^S \leq t_i^S$, $i = 0, 1, \ldots, m$, then complete S by setting $x_j = 0$, $j \in \bar{J}_+^S$ and $x_j = 1$, $j \in \bar{J}_-^S$. Save his solution and its z value as the best feasible solution so far encountered. Terminate examination of completions of S.

Test 2: If $\rho_i^S > t_i^S$ for some i $(i = 0, 1, \ldots, m)$, terminate the examination of completions of S.

Test 3: If $\rho_i^S + |a_{ij}| > t_i^S$ for some i and some $j \in \bar{J}^S$ (say $i = k$ and $j = r$), then terminate the examination of all completions of S in which

a) $x_r = 0$, if $a_{kr} < 0$, or

b) $x_r = 1$, if $a_{kr} \geq 0$.

S is fathomed in Test 1 when its best completion is found to be feasible. In Test 2, S is fathomed when it is determined that (1) none of its completions are feasible or (2) none of its completions are better than the best feasible solution already encountered. Similarly, in Test 3 fathoming occurs for the partial solution, S', formed by augmenting S with $x_r^{S'} = 0$ if $a_{kr} < 0$ or $x_r^{S'} = 1$ if $a_{kr} \geq 0$ $(x_j^{S'} = x_j^S, j \in J^S)$.

II. POSTOPTIMALITY ANALYSIS

The simplicity of the tests in the basic additive algorithm makes it possible, each time a partial solution is fathomed, to *attribute* the partial solution to the particular constraint that caused the fathoming. Thus, for example, if fathoming of a partial solution, S, occurs in Test 2, because for constraint k, $\rho_k^S > t_k^S$, then S is attributable to constraint k. Similarly, in Test 3 if it is constraint k for which $\rho_k^S + |a_{kr}| > t_k^S$, then the augmented partial solution thereby fathomed is attributable to constraint k. If partial solution S is fathomed in Test 1, it may be attributed to constraint 0 (since as the value of $\bar{z}$ continues to decrease, constraint 0 will be violated by the best completion of S).

As we shall now demonstrate, the attributability property affords a useful basis for partitioning the 2^n solutions of the pure (0, 1) problem, one that permits a limited type of ranging analysis on selected parameters and facilitates experimentation with a variety of parameter changes, without the necessity of resolving the entire problem for each one.

Let Λ denote the set of partial solutions fathomed in solving a (0, 1) problem. As long as the enumerative scheme is exhaustive and nonredundant, every solution to the problem will, by definition, be a completion of precisely one member of Λ. It follows therefore that any partition of Λ induces a partition of the 2^n solutions. The procedures outlined here make use of this fact by proceeding in two phases. In the first, a partition of the 2^n solutions is achieved by partitioning the Λ generated in solving a specially constructed (0, 1) problem. This partition is then used in the second phase to undertake the postoptimality analysis.

---

*Several extensions and improvements in the additive algorithm tests have been proposed since their initial statement by Balas [2]. (See Glover and Zionts [10], Peterson [16], Fleischmann [6], Geoffrion [7]). Some of these are reflected here, and others could easily be incorporated into the procedures to be developed. One important device that does require special treatment, the composite constraint, is discussed in section IV.

SOLUTION PHASE:

(a) Formulate and solve (using the additive algorithm) the problem

$$(I^+) \qquad \text{Minimize } z = \underline{c}^+\underline{x} \text{ subject to } \underline{A}^+\underline{x} \leqslant \underline{b}^+, \ x_j = 0 \text{ or } 1,$$

where c_j^+ and a_{ij}^+ are the largest c_j and a_{ij} of interest, respectively, and b_i^+ is the smallest b_i of interest. Observe, however, the following modification of the usual procedure:

(b) Each time a partial solution is fathomed, save the partial solution and attribute it to the appropriate constraint.

At the conclusion of the solution phase, the fathomed partial solutions of (I^+) have been partitioned into $m+1$ disjoint sets. Denote by Λ_k the set of partial solutions attributed to constraint

$$k,\ k = 0, 1, \ldots, m \left(\bigcap_{i=0}^{m} \Lambda_i = \Lambda . \Lambda_i \cap \Lambda_k = \phi,\ i \neq k \right).$$

Problem (I^+) is the tighest (most conservative) formulation of interest. Subsequent parameter changes will involve relaxing constraints or reducing contribution coefficients, implying potentially better solutions. We focus initially on analyzing individual contribution coefficients and individual right-hand sides, but extensions to handle constraint coefficients and multiple parameter changes will follow directly.

We may make several observations concerning the partition achieved in the solution phase:

(a) If the right side of constraint k (b_k) is increased (thereby relaxing the constraint), all other parameters unchanged, then (1) the previous optimal solution is feasible (but not necessarily optimal) in the revised problem, and (2) a solution better than the previous optimum, if it exists, must be a completion of some $S \in \Lambda_k$.

(b) The previous optimal solution will remain optimal for increases in b_k that are less than u_k, where

$$u_k = \min_{S \in \Lambda_k} \{ \rho_k^S - t_k^S \}.$$

(c) If variable x_r is zero in the previous optimal solution and its contribution coefficient (c_r) is reduced, all other parameters unchanged, then (1) the previous optimal solution may not be optimal for the revised problem, but (2) a better solution, if it exists, must be a completion of some $S \in \Lambda_0$ for which x_r is free or $x_r^S = 1$.

(d) If x_r is zero in the previous optimal solution, that solution will remain optimal for decreases in c_r that are less than w_r, where

$$w_r = \min \{ \phi^S \},$$

with

$$\phi^S = \begin{cases} \rho_0^S - t_0^S + c_r, & \text{if } x_r \in \bar{J}_+^S \\ \rho_0^S - t_0^S, & \text{otherwise} \end{cases}$$

and the minimum taken over all $S \in \Lambda_0$ for which x_r is free or $x_r^S = 1$.*

*Since $\bar{z}$ is a quantity that is continually being reduced, the value of t_0^S, when S is reexamined, will in general be smaller than the value when S was fathomed.

(e) If x_r is one at the previous optimum, no reduction in c_r will change the optimal solution (although the objective function will change).

Observation (a) follows from the fact that any $S \in \Lambda_l$, $l \neq k$, has the property that $\rho_l^S > t_l^S$, which is unaffected when parameters in constraint k are revised. Furthermore, $\rho_k^S > t_k^S$ for all $S \in \Lambda_k$, so that solutions better than the current optimal can exist only if parameter changes in constraint k are sufficient to reverse this condition for at least one $S \in \Lambda_k$. This implies observation (b). Similarly, for observations (c), (d), and (e), only partial solutions that are elements of Λ_0 will be affected. We further note, however, that for any partial solution in which x_r is fixed at zero, the property $\rho_0^S > t_0^S$ will be unaffected by a change in c_r.

Ranging Analysis.

Observations (b) and (d) provide the means for doing parameter ranging with respect to the current optimal solution.

POSTOPTIMAL PHASE: *Ranging Analysis* (RA)

(a) *On* b_k: Compute u_k, as defined above. The current best solution will remain optimal for b_k in the interval $b_k^+ - s_k \leq b_k < b_k^+ + u_k$, where s_k is the amount of slack in constraint k at the current optimum.

(b) *On* c_r ($x_r = 0$ at current optimum): Compute w_r, as defined above. The current best solution will remain optimal for any $c_r > c_r^+ - w_r$.

Parameter Changes.

With any partial solution, S, we may associate a *partial problem*, (I^S), of the form

$$(I^S) \qquad \text{Minimize: } z = \sum_{j \in \bar{J}^S} c_j^+ x_j + z^S \qquad \text{subject to } \sum_{j \in \bar{J}^S} a_{ij}^+ x_j \leq b_i^S,\ x_j = 0 \text{ or } 1,\ j \in \bar{J}^S$$

$$\text{where } z^S = \sum_{j \in J^S} c_j^+ x_j^S \qquad \text{and } b_i^S = b_i^+ - \sum_{j \in J^S} a_{ij}^+ x_j^S.$$

Each solution for (I^S), together with S itself, forms a completion of S for (I^+). The best solution for (I^S) will yield the best completion of S for (I^+); and, at the same time, any (in)feasible solution for (I^S) will yield a (in)feasible completion of S in (I^+).

To locate a new optimum when a single parameter is changed, it is simply necessary to solve an appropriate sequence of partial problems.

POSTOPTIMAL PHASE: *Single Parameter Changes* (SPC)

(a) *Revision of* b_k: For each $S \in \Lambda_k$ formulate (I^S) using the revised value of b_k and an initial value of $\bar{z}$ determined by the best feasible solution so far encountered (in the solution phase or in any partial problem previously solved). Solve this problem completely using the additive algorithm. When every (I^S), $S \in \Lambda_k$, has been examined, a new optimal solution, if it exists, will have been found.

(b) *Revision of* c_r: For each $S \in \Lambda_0$ in which x_r is not fixed at zero, formulate (I^S) using the revised value of c_r and an initial value of $\bar{z}$ determined by the best feasible solution so far encountered. Solve this problem completely using the additive algorithm. When every such (I^S) has been examined, a new optimal solution, if it exists, will have been found.*

*In section III there is a discussion of the implications of attributing and saving partial solutions or not doing so as problems are solved in the postoptimal phase.

By this process only selected solutions of problem (I^+) are reexamined in the postoptimal phase, thereby making it feasible to study a variety of parameter changes without the relatively heavy computational burden associated with solving and re-solving a (0, 1) problem.*

III. FURTHER COMMENTS

Multiple Parameter Changes

The procedures for treating multiple parameter changes are a straightforward generalization of the single parameter procedures outlined in the previous section.

POSTOPTIMAL PHASE: *Multiple Parameter Changes* (MPC)

(a) *Revision of* b_i, b_k, . . ., *and* b_l: For each $S\epsilon\Lambda_i \cup \Lambda_k \cup \ldots \cup \Lambda_l$ formulate (I^S) using the revised values of b_i, b_k, . . ., and b_l and an initial value of $\bar{z}$ determined by the best feasible solution so far encountered. When every (I^S), $S\epsilon\Lambda_i \cup \Lambda_k \cup \ldots \cup \Lambda_l$, has been examined, a new optimum, if it exists, will have been found.

(b) *Revision of* c_q, c_r, . . ., *and* c_s: For each $S\epsilon\Lambda_0$ in which x_q, x_r, . . ., and/or x_s is not fixed at zero, formulate (I^S) using the revised values of c_q, c_r, . . ., and c_s and an initial value of $\bar{z}$ determined by the best feasible solution so far encountered. When every such (I^S) has been examined, a new optimum, if it exists, will have been found.

Constraint Coefficients.

The procedures for analyzing constraint coefficients are very similar to those for contribution coefficients. With respect to ranging analysis, we may define

$$v_{kr} = \min\{\Psi^S\},$$

where

$$\Psi^S = \begin{cases} \rho_k^S - t_k^S + a_{kr}, & \text{if } x_r \text{ is free and } a_{kr} \geq 0 \\ \rho_k^S - t_k^S, & \text{otherwise} \end{cases}$$

and the minimum is taken over all $S\epsilon\Lambda_k$ in which x_r is free or $x_r^S = 1$. Then we may add to *RA* above

(c) *On* a_{kr}: Compute v_{kr}, as defined above. The current best solution will remain optimal for a_{kr} in the interval $a_{kr}^+ - v_{kr} < a_{kr} \leq a_{kr}^+ + s_k$, where s_k is the amount of slack in constraint k at the current optimum. For the case where multiple constraint coefficients are revised, we may add to MPC above

(c) *Revision of* a_{iq}, a_{kr}, . . ., *and* a_{ls}: For each $S\epsilon\Lambda_i \cup \Lambda_k \cup \ldots \Lambda_l$ in which $x_q, x_r, \ldots$ and/or x_s is not fixed at zero, formulate (I^S) using the revised values of a_{iq}, a_{kr}, . . ., and a_{ls} and an initial value of $\bar{z}$ determined by the best feasible solution so far encountered. When every such (I^S) has been examined, a new optimum, if it exists, will have been found.

Deletion of Constraints.

The deletion of constraints i, k, . . ., and l may be handled with procedure MPC(a) by simply using arbitrarily large values for b_i, b_k, . . ., and b_l (or, equivalently, deleting constraints i, k, . . ., and l from each (I^S) as it is formulated).

*The handling of equality constraints in this process requires special comment. It is necessary to view the constraint $\sum_{j=1}^{n} a_j x_j = b$ in (I^+) as two constraints, $l \leq \sum_{j=1}^{n} a_j x_j \leq u$, where initially $l = u = b$. Then, "relaxing" the constraint takes the form of either decreasing l or increasing u. It is generally not possible to simply change the value of b in the post-optimal phase, since this would require relaxing one of the constraints but tightening the other in a corresponding way.

Addition of Variables.

The addition of variables q, r, . . ., and s is most easily handled using the somewhat artificial device of including in problem (I^+) *every* variable of interest, but initially assigning arbitrarily large values to c_q, c_r, . . ., and c_s. Then "addition" of these variables takes the form of reducing the values of c_q, c_r, . . . and c_s and using procedure MPC(b).*

Permanent Parameter Changes.

The parameters of (I^+), together with the additive algorithm itself, determine the partition of solutions achieved in the solution phase. In the postoptimal phase, changes in these parameters may be rendered permanent or temporary (at the user's discretion), depending on whether or not the initial partition is appropriately revised to reflect them. This revision may be accomplished, while solving the required sequence of partial problems, by (1) removing each partial solution from its subset as it is examined and (2) solving the associated partial problems with the additive algorithm as modified in the solution phase. If this is not done, then every new parameter change is, in effect, made in problem (I^+), and it is in this sense that the change is temporary.

IV. COMPOSITE CONSTRAINTS

One device for strengthening the additive algorithm is that of appending one or more *composite* or *surrogate* constraints to the original problem (see Balas [3], Geoffrion [8], and Glover [9]). In general, the number of composite constraints employed and their form will be a function of the current state of the search. When partial solution S is under examination, a composite constraint is generated by taking a positive linear combination of the constraints of (I^S):†

$$\sum_{i=1}^{m} y_i \sum_{j \in \bar{J}^S} a_{ij} x_j \leq \sum_{i=1}^{m} y_i b_i^S .$$

A "strongest" composite constraint may be computed by choosing the values of $y_1, \ldots, y_m$ to be the dual variables of the linear program formed from (I^S) by relaxing "$x_j = 0$ or $1, j \in \bar{J}^S$" to "$0 \leq x_j \leq 1$, $j \in \bar{J}^S$."

When composite constraints are employed, certain modifications of the earlier procedures are required. In the solution phase, a composite constraint may be treated as any other constraint is treated, provided that the weights used to generate it are noted. But in the postoptimal phase, when a change is made in, say, constraint k, or constraint k is deleted, it is necessary to examine all partial solutions attributed to composite constraints for which y_k is nonzero, as well as those attributed directly to constraint k. The procedures outlined in sections II and III may be modified easily to reflect this fact.**

In fact, the solution of the linear programming problem formed from (I^S) (or the solution of its dual) has value beyond its usefulness in computing a composite constraint:

---

*Alternatively, these variables need not be initially included. In this case, for every k such that $a_{kq} < 0$ and/or $a_{kr} < 0$, . . ., and/or $a_{ks} < 0$, all $S \in \Lambda_k$ must be examined.

†Constraint 0 is sometimes included in the linear combination as well (Geoffrion [8]).

**Also, as each partial solution is fathomed in the solution phase (by any constraint in the problem), it is useful to make provision for "recapturing" in the postoptimal phase the composite constraints that were in use when fathoming occurred. Such constraints can be made a permanent part of the associated partial problem, if it should be utilized in the second phase.

(1) The optimal linear programming solution provides a lower bound on the value of every completion of S. This lower bound may, on occasion, be above the current $\bar{z}$.

(2) If the optimal linear programming solution should happen to be all-integer, it would necessarily constitute a best feasible completion of S.

(3) Infeasibility in the relaxed problem, whenever it is encountered, will imply infeasibility in (I^S).

Since all components of the problem are simultaneously under examination here, the partial solution fathomed in any of these three ways cannot be attributed to a single constraint. It becomes necessary, therefore, to build a set of *unattributed* partial solutions in the solution phase, to be examined in the postoptimal phase with every parameter change.*

V. IMPLEMENTATION

The procedures outlined here have been programmed, with minor computational modifications, for the GE635 time-sharing computer. As with most methods of postoptimality analysis, implementing the procedures in a time-sharing mode has greatly enhanced their effective utilization. Once the problem has been entered and an initial optimal solution obtained, the user has the option of operating in a variety of ways. He may, for example, wish to track the optimal solution over a sequence of permanent parameter changes and test the sensitivity of each using ranging analysis. Alternatively, he may simply want to compare the optima generated over a set of temporary parameter changes. The program has been used, for example, to experiment with alternative borrowing plans in several capital budgeting problems (which involves manipulating right hand sides).

The computational burden at the postoptimal stage has generally been found to be quite low.† In one capital budgeting problem involving 20 projects (variables) over 10 periods (constraints), a total of 1,091 partial solutions were examined in the solution phase (using an L.P. initial solution) and an additional 1,143 partial solutions were examined in studying five other borrowing plans. This is in contrast with the 6,515 partial solutions that required examination to solve the same five problems separately, using the original optimum as an initial solution in each case. There has been similar experience with every other problem tested for all but the most drastic parameter changes (and even in these cases, the results were no worse than they would have been if the problem had been reformulated and solved from the beginning). Conclusive computational experience with the methods in this paper, however, remains to be accumulated.

APPENDIX

AN EXAMPLE

Consider the problem:

$$\text{Minimize: } z = 2x_1 + 5x_2 + 5x_3 + 6x_4 + 4x_5 + x_6 + 8x_7 + x_8$$

---

* Unattributed partial solutions can, nevertheless, be analyzed in ways useful for postoptimality analysis. This topic is the focus of work currently underway.

† In the current version of the program, the time required to attribute and save fathomed partial solutions in random access files has generally been 25 to 30 percent of total computation time in the solution phase.

Subject to:

$$-2x_1 + 2x_4 - 6x_5 + x_6 - x_7 + 2x_8 \leq -5$$

$$-4x_1 + 11x_2 - 11x_3 - 7x_4 + 4x_5 + 3x_6 - 5x_7 + x_8 \leq -6$$

(A1)

$$x_2 + x_3 + x_4 - x_5 - 2x_6 + x_8 \leq 0$$

$$x_j = 0 \text{ or } 1, \qquad j = 1, \ldots, 8$$

The solution of this problem is now given in some detail to demonstrate the procedures previously presented. We will utilize only the tests outlined in section 1. The organization of the search will be essentially that proposed by Geoffrion [7]. The partial solution, S, is maintained as a vector in which "j" ("$-j$") is an element if variable j is fixed at 1 (0), and the elements are ordered in the vector as they are generated. Underlining an element "j" commemorates the fathoming of the partial solution defined by "$-j$" and the elements that precede it in the vector. The augmentation rule used is the same one proposed by Balas [2]: Augment S with the $r \epsilon \bar{J}^S$ which maximizes $\sum_{i=0}^{3} \min \{t_i - a_{ij}, 0\}$ over all $j \epsilon \bar{J}^S$. Initially, we take $\bar{z} = 999$.

Solution Phase.

The sequence of partial solutions examined in solving problem (A1) appear in the first column of Table I. In the second column the associated vectors, $\underline{\rho}^S = (\underline{\rho}_0^S, \ldots, \rho_3^S)$ and $\underline{t}^S = (t_0^S, \ldots, t_3^S)$, appear. (Since $c_j > 0$, $j = 1, \ldots, 8$ in this problem, $\delta^S = (\delta_0^S, \ldots, \delta_3^S)$ will always be a vector of zeroes. It is therefore not shown in Table I). Each time a partial solution is fathomed, it is assigned the number appearing in the last column of Table I and entered in Table II. Initially $\underline{\rho} = (0, -9, -27, -3)$ and $t = (999, -5, -6, 0)$.

The optimal solution to problem (A1) is $x_3 = x_5 = 1$, all other variables equal zero, $z = 9$.

Postoptimal Phase.

In Table II each fathomed partial solution has been attributed to the appropriate constraint by placement in one of four cells. In brackets following the partial solutions in cell i ($i = 1,2,3$) are the quantities $(\rho_i^S - t_i^S)$ and $\rho_0^S - (t_0^S - \bar{z})$. The first of these is the amount by which the completion of S nearest to being feasible in constraint i violates it; the second is the value of the best completion of S. Only the second of these two quantities follows the partial solutions in cell 0.

Ranging analysis on the contribution coefficients requires examination of the entries in cell 0. Since x_1 is not at 1 in the current optimum and is not fixed at 0 by any entry in cell 0, a reduction in c_1 by an amount less than $w_1 = \min \{(18 - 9 + 2), (17 - 9 + 2), (9 - 9 + 2), (11 - 9), (12 - 9)\} = 2$ will leave the current optimum unchanged. For x_2, which is fixed at zero in partial solutions 5 and 10, $w_2 = \min \{(18 - 9 + 5), (9 - 9 + 5), (11 - 9)\} = 2$. Similarly, $w_4 = 3$, $w_6 = 1$, $w_7 = 8$, and $w_8 = 1$. Since x_3 and x_5 are at 1 in the current optimum, we may take $w_3 = w_5 = 999$. Thus, $(w_1, \ldots, w_8) = (2, 2, 999, 3, 999, 1, 8, 1)$.

For ranging on the right side of constraint i, we examine the partial solutions in cell i. Thus, a change in b_1 by an amount less than $u_1 = \min \{2, 3\} = 2$ will not admit the possibility of a better feasible solution. In computing u_2, partial solutions 3 and 4 may safely be ignored, since the value

TABLE I

Partial solution	Status vectors	Comments	
$\{7\}$	$\underline{\rho} = (0, -8, -22, -3)$ $\underline{t} = (991, -4, -1, 0)$	Forward Step	
$\{7, \underline{5}\}$	$\underline{\rho} = (0, -2, -22, -2)$ $\underline{t} = (987, 2, -5, 1)$	Variable 5 Fixed at 1 by Constraint 1	(1)
$\{7, \underline{5}, 4\}$	$\underline{\rho} = (0, -2, -15, -2)$ $\underline{t} = (981, 0, 2, 0)$	Forward Step	
$\{7, \underline{5}, 4\}$	$\underline{\rho} = (0, -2, -15, -2)$ $\underline{t} = (0, 0, 2, 0)$	Better Feasible Solution	(2)
$\{7, \underline{5}, -\underline{4}\}$	$\underline{\rho} = (0, -2, -15, -2)$ $\underline{t} = (6, 2, -5, 1)$	Backtrack	
$\{7, \underline{5}, -\underline{4}, -\underline{2}\}$	$\underline{\rho} = (0, -2, -15, -2)$ $\underline{t} = (6, 2, -5, 1)$	Variable 2 Fixed at 0 by Constraint 2	(3)
$\{7, \underline{5}, -\underline{4}, -\underline{2}, \underline{3}\}$	$\underline{\rho} = (0, -2, -4, -2)$ $\underline{t} = (1, 2, 6, 0)$	Variable 5 Fixed at 1 by Constraint 1	(4)
$\{7\ \underline{5}, -\underline{4}, -\underline{2}, \underline{3}\}$	$\underline{\rho} = (0, -2, -4, -2)$ $\underline{t} = (0, 2, 6, 0)$	Better Feasible Solution	(5)
$\{-\underline{7}\}$	$\underline{\rho} = (0, -8, -22, -3)$ $\underline{t} = (17, -5, -6, 0)$	Backtrack	
$\{-\underline{7}, \underline{5}\}$	$\underline{\rho} = (0, -2, -22, -2)$ $\underline{t} = (13, 1, -10, 1)$	Variable 5 Fixed at 1 by Constraint 1	(6)
$\{-\underline{7}, \underline{5}, 3\}$	$\underline{\rho} = (0, -2, -11, -2)$ $\underline{t} = (8, 1, 1, 0)$	Forward Step	
$\{-\underline{7}, \underline{5}, 3\}$	$\underline{\rho} = (0, -2, -11, -2)$ $\underline{t} = (0, 1, 1, 0)$	Better Feasible Solution	(7)
$\{-\underline{7}, \underline{5}, -\underline{3}\}$	$\underline{\rho} = (0, -2, -11, -2)$ $\underline{t} = (5, 1, -10, 1)$	Backtrack	
$\{-\underline{7}, \underline{5}, -\underline{3}, \underline{1}\}$	$\underline{\rho} = (0, 0, -7, -2)$ $\underline{t} = (3, 3, -6, 1)$	Variable 1 Fixed at 1 by Constaint 2	(8)
$\{-\underline{7}, \underline{5}, -\underline{3}, \underline{1}, -\underline{2}\}$	$\underline{\rho} = (0, 0, -7, -2)$ $\underline{t} = (3, 3, -6, 1)$	Variable 2 Fixed at 0 by Constraint 0	(9)
$\{-\underline{7}, \underline{5}, -\underline{3}, \underline{1}, -\underline{2}, -\underline{4}\}$	$\underline{\rho} = (0, 0, 0, -2)$ $\underline{t} = (3, 3, -6, 1)$	Variable 4 Fixed at 0 by Constraint 0	(10)
$\{-\underline{7}, \underline{5}, -\underline{3}, \underline{1}, -\underline{2}, -\underline{4}\}$	$\underline{\rho} = (0, 0, 0, -2)$ $\underline{t} = (3, 3, -6, 1)$	Infeasibility in Constraint 2	(11)

TABLE II

Constraint 0			Constraint 1		
(2)	{7, 5, 4}	[18]	(1)	{7, −5}	[2, 8]
(5)	{7, 5, −4, −2, 3}	[17]	(6)	{−7, −5}	[3, 0]
(7)	{−7, 5, 3}	[9]			
(9)	{−7, 5, −3, 1, 2}	[11]			
(10)	{−7, 5, −3, 1, −2, 4}	[12]			
Constraint 2			**Constraint 3**		
(3)	{7, 5, −4, 2}	[1, 17]			
(4)	{7, 5, −4, −2, −3}	[1, 12]			
(8)	{−7, 5, −3, −1}	[3, 4]			
(11)	{−7, 5, −3, 1, −2, −4}	[6, 6]			

of the best completion in each is poorer than the value of the current best solution. Hence $u_2 = \min\{3, 6\} = 3$. Since no partial solutions are attributed to constraint 3, we may take $u_3 = 999$. Thus, $(u_1, u_2, u_3) = (2,3,999)$.

Suppose now that c_2 and c_4 are both reduced to 2. The entries in cell 0 must again be considered. Since partial solution 5 has both x_2 and x_4 fixed at 0, it may be disregarded. Partial solutions 2 and 7 may also be disregarded since in neither one are the proposed parameter changes sufficient to move the value of the best completion (18 or 9) below the current value of $\bar{z}$ (9). Thus, two partial problems, those formulated from partial solutions 9 and 10, must be solved:

(9): Minimize: $z = 6x_4 + x_6 + x_8 + 7$

Subject to:

$$2x_4 + x_6 + 2x_8 \leq 3$$

$$-7x_4 + 3x_6 + x_8 \leq -17$$

$$x_4 - 2x_6 + x_8 \leq 0$$

$$x_4, x_6, x_8 = 0 \text{ or } 1$$

(10): Minimize: $z = x_6 + x_8 + 8$

Subject to:

$$x_6 + 2x_8 \leq 1$$

$$3x_6 + x_8 \leq 1$$

$$-2x_6 + x_8 \leq 0 \qquad [14]$$

$$x_6, x_8 = 0 \text{ or } 1 \qquad [15]$$

The first of these has no feasible solution. In the second, $x_6 = x_8 = 0$ is feasible, so that a better solution, $x_1 = x_4 = x_5 = .1, z = 8$, is located. [16]

If, instead, b_1 were increased to -2, the partial problems associated with partial solutions 1 and 6 would require formulation and solution. One may verify, however, that neither of these problems leads to a feasible solution that is better than the original optimum.

Finally, suppose that constraint 2 were dropped from the problem. Partial solutions 8 and 11 would require consideration. The partial problem for the first of these yields a better feasible solution, $x_5 =$ $z = 4$. Consequently, the second problem need not be formulated, since the value of its best completion (6) exceeds the value of the current best feasible solution.

BIBLIOGRAPHY

[1] Alcaly, R. F. and A. K. Klevorick, "A Note on the Dual Prices of Integer Programs," Econometrica *34*, 206–214 (1966).

[2] Balas, E., "An Additive Algorithm for Solving Linear Programs with Zero-One Variables," Operations Research *13*, 517–546 (1965).

[3] Balas, E., "Discrete Programming by the Filter Method," Operations Research *15*, 915–957 (1967).

[4] Balas, E., "Duality in Discrete Programming IV. Applications," M.S.R.R. No. 145, Carnegie-Mellon University (1968).

[5] Balinski, M. L. and K. Spielberg, "Methods of Integer Programming: Algebraic, Combinatorial, and Enumerative," in *Progress in Operations Research*, edited by J. S. Aronofsky (Wiley and Sons, New York, 1969).

[6] Fleischmann, B., "Computational Experience with Algorithm of Balas," Operations Research *15*, 153–155 (1967).

[7] Geoffrion, A. M., "Integer Programming by Implicit Enumeration and Balas Method," SIAM Review *9*, 178–190 (1967).

[8] Geoffrion, A. M., "An Improved Implicit Enumeration Approach For Integer Programming," Operations Research *17*, 437–454 (1969).

[9] Glover, F., "A Multiphase-Dual Algorithm for the Zero-One Integer Programming Problem," Operations Research *13*, 879–919 (1965).

[10] Glover, F. and S. Zionts, "A Note on the Additive Algorithm of Balas," Operations Research *13*, 546–549 (1965).

[11] Gomory, R. E., and W. J. Baumol, "Integer Programming and Pricing," Econometrica *28*, 521–550 (1960).

[12] Jensen, R. E., "Sensitivity Analysis and Integer Linear Programming," The Accounting Review *43*, 425–446 (1968).

[13] Land, A. H., and A. G. Doig, "An Automatic Method of Solving Discrete Programming Problems," Econometrica *28*, 497–520 (1960).

[14] Lemke, C. E. and K. Spielberg, "Direct Search Zero-One and Mixed Integer Programming," Operations Research *15*, 892–914 (1967).

[15] Little, J. D. C., K. G. Murty, D. W. Sweeney, and C. Karel, "An Algorithm for the Travelling Salesman Problem," Operations Research *11*, 972–989 (1963).

[16] Peterson, C. C., "Computational Experience with Variants of the Balas Algorithm Applied to the Selection of R and D Projects," Management Science *13*, 736–750 (1967).

STATISTICAL AUDITING OF LARGE-SCALE MANAGEMENT INFORMATION SYSTEMS

Sheldon E. Haber, Frank W. Segel, and Henry Solomon

Department of Economics
and
Institute for Management Science and Engineering
The George Washington University

ABSTRACT

In recent years many large government and commercial organizations have developed large-scale management information systems. During the early phases of any system development, questions are likely to be raised concerning the criteria of completeness and accuracy of the data. How to respond to these questions is difficult, and represents the general subject of this paper. In particular, this paper describes the application of statistical techniques, namely, various kinds of multivariate data analysis, to the problem of assessing the quality of a large-scale data collection system in the U.S. Navy. Results of using these techniques are described and indicate their usefulness as auditing procedures. While the discussion is placed in a particular context, the procedures should be of value for other large-scale information systems.

INTRODUCTION

In recent years many large government and commercial organizations have been developing large-scale management information systems. The general direction has been to make these systems as comprehensive as possible. That is, to satisfy simultaneously the many interests and levels of management within the organization. Usually a large-scale data collection system is implemented incorporating many detailed data elements. While the form of these data may vary, the data are typically transmitted from source reporting activities to some central data processing facility. These data are then transformed into computer master and/or work files and used as inputs to algorithms relevant to the problems of the particular organization.

During the early phases of any system development, questions are likely to be raised concerning the criteria of completeness and accuracy of the data. How to respond to these questions is difficult, and represents the general subject of this paper. The difficulties are due to various factors. One is the large volume of data. The system representing the particular context of this paper is currently generating close to three million EAM cards per month. Obtaining subjective judgments or objective appraisals of completeness and accuracy of these huge amounts of data creates problems of a different kind rather than only a difference in degree in comparison to data collection systems of more modest and traditional size.* A second factor is the large range of data elements. As more data elements are added, the task of appraising completeness and accuracy becomes increasingly difficult. But perhaps the most important obstacle in appraising the data is the usual lack of available standards. The problem

*This and other factors suggest an important problem which is not dealt with in this paper, i.e., the economies and diseconomies of scale of management information systems. The general direction of creating a single comprehensive system deserves more scrutiny than it has received.

to be faced is that the information system is intended to supply information needed to establish standards for assessing quality of data without prior knowledge of the quality of the initial information upon which such standards are to be based.

The remainder of this paper is a description of the application of multivariate data analysis for assessing the quality of a large-scale data collection system in the U.S. Navy. The results of using this approach suggest a useful means for tackling the difficulties mentioned above. It should be noted that the thrust of the paper is directed toward the application of statistical techniques for auditing large-scale management information systems rather than analysis of the statistical techniques themselves. While the discussion is placed in a particular context, the procedures should be of value for other large-scale information systems.

THE U.S. NAVY MAINTENANCE DATA COLLECTION SYSTEM

A major management information system in the Navy is the Maintenance and Material Management Information System, often referred to as the 3M Program. The System includes an extremely large-scale data collection subsystem pertaining to the maintenance process. The specific context of this paper is this subsystem termed the "Maintenance Data Collection System" (MDCS).

Large volumes of data are being collected from the air and surface fleets, which define in considerable detail the nature and extent of maintenance activity being conducted by organizational and intermediate echelons of maintenance support and their use of resources.* It is anticipated that in the future, major policy and allocation decisions will be based on analyses and operational use of these data. Hence it is important that the data base is acceptable with respect to its accuracy and completeness.

Some aspects of the auditing process are fairly straightforward. A source data document can be considered to be identified as containing errors if it does not conform with prescribed standards pertaining to form. In this context, one type of error is an inadmissible data element. A simple example of this error is a field which should contain a certain number of alpha positions but none are observed.

Another kind of error which may be handled by a straightforward but more extensive procedure is one which requires the use of master data files and associated data processing. An illustration of this is the verification of a repair part stock number on a source document which conforms to format requirements as mentioned above. The stock number may still be identified as an error if it is determined via the use of a master repair parts file that the number does not exist in the supply system.

There are other types of errors which are much more difficult to identify. These pertain to measurement errors, where measurement may relate, for example, to the amount of man-hours reported for a job. The principal difficulty of error identification stems from the relative lack of knowledge of the maintenance process. This, of course, is what the 3M System is intended to correct. However, presently there is what may be characterized as a circular problem. If more were known about maintenance activities, criteria may be determined for evaluating measurement errors, but the MDCS is itself intended to remedy this lack of knowledge.

The approach taken in this paper to distinguish between good vs poor data is to identify good vs poor reporters. More specifically, the goal is the identification of under-reporting and the general quality of source reporters without an exhaustive search and critique of individual data forms and data elements. Put another way, in this paper the problem of auditing large volumes of data is ex-

*Organizational maintenance refers to work done by the ship's force. Intermediate maintenance refers to work done for a ship by tenders and repair ships.

amined from the point of view of assessing the overall quality of data rather than the quality of individual data elements. The choice of this problem is dictated by taking first things first. Again, this task must generally be performed in the absence of *a priori* standards. The capability for determining good vs poor reporters under this constraint represents the thrust of the methodology and results to be described.

DATA BASE FOR ANALYSIS

In order to pursue efficiently the feasibility of distinguishing between good vs poor reporters, the data base was restricted to the experience of 62 destroyers from a total population of 140 for an 18-month period.* Given the sample ships, a stratification was accomplished in terms of organizational maintenance card counts. By this stratification, three groups of ships denoted as A, B, and C were formed to distinguish "good" and "bad" reporters. Numerical criteria for forming these groups are given in Table I.

TABLE I

Group	Number of ships		Average monthly card count per ship
	Total population	Sample	
A	18	18	725 or more
B	75	30	500–724
C	47	14	Less than 500

The rationale for the stratification is extremely important. Since ships may vary with respect to their quality as data sources, a basic proposition underlying the study is that differences in card counts may reflect differences in quality. This inference was used as a basis for stratifying the population and selecting the sample.

If one of a few simple measurements, such as card counts, can be used to differentiate between good vs poor reporters, mechanized auditing procedures become feasible and acceptable. It was recognized, however, that the problem of defining and identifying good vs poor reporters is complicated by the determinants of the maintenance process itself. It may be assumed that the amount of data submitted by a ship is dependent on the level and kinds of maintenance performed, administrative practices in carrying out maintenance policies, operational activity levels, etc. To the extent there are differences between ships or groups of ships due to these factors, one may expect differences in volume of reporting which are not due to differences in quality of reporting. Hence there must be some attempt to isolate the many variables that influence volume of reporting in order to estimate their importance in determining which ships are good and which are poor reporters.

APPLICATION OF REGRESSION ANALYSIS

Multiple regression analysis was used to isolate variables which might account for variations in reporting and which may have little to do with the quality of data reporting. Thus, the principal motivation for this approach in the present context is to determine the presence of structural relation-

*If available data from all ships were used, this would have involved 13 to 14 million surface fleet records requiring an infeasible amount of computer processing. For the sample considered here, there were three-quarters of a million transaction cards.

ships where, for example, these pertain to relationships between card counts and maintenance and operational activity variables. Of secondary importance, but still of some interest, is the extent to which information pertaining to these activity variables can be used to explain deviations in card volume from expected levels.

Two sets of computations were performed. All 62 ships were included in the first set. The second set involved the partitioning of the 62 ships into their respective groups (i.e., A, B, and C). Separate computations were then performed utilizing the data for each group.

Although attention is focused on the relationship between total card count for the 18-month period as the dependent variable and a set of independent variables, computations were performed utilizing other dependent variables.* The independent variables included: "age of ship," "underway time," "time in MDCS," "light tonnage," "combat deployment," "total system months in preventive maintenance system (PMS)," "time in overhaul," and "level of intermediate maintenance activity."

The age of ship is self-explanatory. Underway time is the amount of time during the 18-month interval that the ship was underway and represents an important dimension of operational activity. The time in MDCS is the total number of months the ship was reporting under MDCS as of the end of the 18-month period and represents an experience factor in reporting of maintenance data. Light tonnage is also self-explanatory and serves to distinguish the possible relationship between size of ship and level of maintenance activity. Combat deployment refers to the number of days the ship was in a combat zone, reflecting another dimension of operational activity. Total system-months in PMS serves to distinguish the effect of this type of maintenance activity and differences in PMS implementation dates for the various shipboard systems. Days in overhaul is intended to distinguish the amount of time each ship spent in overhaul diring the 18-month period and its possible diffcrential effects on organizational and intermediate level maintenance. Intcrmediate maintenance activity is approximated by an index based on the ratio of intermediate maintenance man-hours to the total number of intermediate and organizational maintenance man-hours. This variable is intended to reflect differences among ships in reliance on intermediate maintenance support.

As noted above, the first set of computations included all 162 ships without refcrence to their card count group. The results of this computation are shown below in Table II.

As may be noticed, no significant relationship was found between card volume and age of ship, underway time and tonnage or size of ship. The regression coefficients for the remaining four variables were found to be significantly different from zero. The positive regression coefficient for time in MDCS and the negative coefficients for the overhaul and intermediate maintenance variables are not surprising. The negative coefficient for the combat deployment variable and the only one which is significant at the 0.01 level is, however, contrary to the prior expectation of a positive relationship.

A second observation is that although the coefficient of determination is significant, it is not very large. The use of the regression equation still leaves approximately one-half of the variation in card volume unexplained. Given the large number of ships with low card volumes in Table I, this is particularly important. In order to assess this aspect of the auditing problem, a second set of regressions was performed. Separate regressions were computed for each of the three ship groups, A, B, and C, to determine whether or not the regression coefficients, found significant when all ships were considered, would likewise be significant within ship groups.

*The other dependent variables included "total maintenance man-hours" and "total number of maintenance jobs." General results of using these measurements as dependent variables are the same as in the case of "card counts." This is not surprising since one may expect a high correlation between card counts, reported man-hours, and the reported number of jobs.

TABLE II. *Sixty-Two Ships*

Variables[a]	Regression coefficient	Standard error
2	−1.214	4.149
3	0.395	0.449
4	283.885[b]	143.103
5	1.981	1.736
6	−17.734[c]	4.106
7	−17.044[b]	7.030
8	−7.201[b]	3.154
	$R^2 = 0.54$[c]	

[a] Variables:
1—Total card count (dependent variable)
2—Age of ship (months)
3—Underway time (hours)
4—Time in MDCS (months)
5—Light tonnage
6—Combat deployment (days)
7—Overhaul (days)
8—Index of intermediate maintenance activity level
[b] Significant at 0.05 level.
[c] Significant at 0.01 level.

TABLE III

Variables[a]	Group A		Group B		Group C	
	Regression coefficient	Standard error	Regression coefficient	Standard error	Regression coefficient	Standard error
2	1.433	34.040	−3.264	2.903	−6.614	40.456
3	0.277	0.785	−1.258	0.668	−0.409	0.502
4	36.480	367.588	331.575[c]	131.517	−44.709	221.908
5	1.481	7.853	−0.968	1.761	−8.596	5.318
6	−4.784	9.778	1.703	5.482	−8.754[b]	4.450
7	−4.622	17.449	−19.787[c]	7.416	−9.563	9.870
8	−12.491	8.149	4.795	2.676	−2.687	6.957
$R^2 = 0.30$			$R^2 = 0.53$[b]		$R^2 = 0.74$	

[a] Variables:
1—Total card count (dependent variable)
2—Age of ship (months)
3—Underway time (hours)
4—Time in MDCS (months)
5—Light tonnage
6—Combat deployment (days)
7—Overhaul (days)
8—Index of intermediate maintenance activity level
[b] Significant at 0.05 level
[c] Significant at 0.01 level

Total card count was the dependent variable: the independent variables were those just described. Numerical results are shown in Table III.

The most interesting finding is the reduction in the number of variables with a significant re-

gression coefficient. For example, the coefficient for intermediate maintenance was significant in the first regression, but was not significant in any of the regressions by ship group. Additionally, the other three variables significant in the first regression were significant in only one case each in the second set of regressions.

The second finding of interest is that the coefficient of determination is significant for only one of the three ship groups. Of note, too, is that in most cases the size of the coefficients was reduced, particularly for the combat deployment variable. Thus, it is seen that the structural relationships between card volume and maintenance and operational activity variables, evidenced in terms of the first regression, is much weaker when the three groups of ships are examined separately. We conclude that while activity variables explain a portion of the variation in card volumes, they do not provide an adequate explanation for the low levels of reporting, particularly for ships in group B and C.

The finding of structural relationships between card volume and maintenance and operational activity variables when the 62 ships are grouped together and a much weaker relationship when the ships are separated into groups is worth noting. It points up an important problem of aggregation in many contexts where regression analysis is applied, and suggests that structural relationships based on aggregated data require further examination to determine their consistency within subsets of observations.*

APPLICATION OF DISCRIMINANT ANALYSIS

Thus far it has been indicated that differences in card counts could not be explained adequately by variations in variables which might be thought to influence maintenance activity levels such as underway steaming hours, time in overhaul, utilization of intermediate maintenance services, etc. One may ask, are there alternative ways, besides the original *ad hoc* three-way classification of ships based on card count, of grouping ships in terms of their adherence to the rules for reporting maintenance activity? One such procedure is discriminant analysis which has its origin in the problem of classification (see [1] and [3]). In taking this tack, the auditing problem is viewed as one of categorizing data reporters.

In the present context, measurements are taken pertaining to the quality of reporting by a given ship, and it is desired to classify the ship on the basis of these measurements into different reporting categories.† In general, if there is a high probability that an object (e.g., a ship) can be correctly classi-

---

*The conclusion of only a very weak structural relationship between card volumes and activity variables was further corroborated by the application of cluster analysis. In cluster analysis, a classification system, such as the grouping of ships into groups A, B, and C, is not determined *a priori*, but instead is determined by the data. The cluster procedure used was developed by King [2]. Using this procedure it was found that differences in card counts among ships are, at best, only partially related to time in MDCS, combat deployment, time in overhaul and intermediate maintenance, i.e., the four variables with significant regression coefficients in Table II. A more detailed description of this application of cluster analysis may be obtained from the authors upon request.

†In discriminant analysis, linear functions of the measurements are used. If the measurements are jointly normally distributed within each reporting category, with the same covariance matrices, the use of linear functions for classification maximizes the probabilities of correct classification. For any two reporting categories, the linear discriminant function

$$Z = a_1x_1 + a_2x_2 + \ldots + a_nx_n,$$

maximizes the differences in the mean values of the functions for the two categories relative to the variations of the values within the categories. The general objective is to find that linear combination of the various measurements which will best discriminate between the two categories.

fied on the basis of measurements, the measurements can then replace the standards used in distinguishing categories in the first place. This is especially convenient when application of the standards is expensive or difficult in contrast to the cost or difficulty of obtaining the measurements used in the discriminate analysis.

The measurements selected for the discriminant analysis pertained to quality of reporting, where quality refers to those aspects of good reporting independent of the quantity or volume of data submissions. The rationale for this choice of measurements was twofold. The first objective was to determine whether it was possible to arrive at a set of qualitative measurements which could be used for ranking ships in terms of responsiveness to instructions in the MDCS program. The second objective was to determine whether qualitative differences in reporting were associated with quantitative differences. To the extent that such an association exists, it would confirm the conclusion reached in the previous section that for the data system examined, low card volume is a real rather than apparent indicator of poor reporting.

TABLE IV

Variable	Mean			Standard deviation		
	A	B	C	A	B	C
x_1: Percent of fields in error	0.1	0.1	0.2	0.1	0.2	0.3
x_2: Percent of cards in error	1.1	1.5	1.9	0.9	2.1	1.9
x_3: Number of months required to receive 90 percent of cards	2.3	2.8	3.3	0.5	0.4	0.7
x_4: Percent of "delay" jobs completely documented	48.9	38.0	28.4	10.3	15.3	16.7
x_5: Percent of "parts" jobs completely documented	41.2	37.4	35.1	2.7	5.3	5.7
x_6: Percent of parts cards with no associated maintenance action card	19.8	31.0	33.9	10.1	11.0	12.9
x_7: Percent of "work request" jobs completely documented	36.1	25.3	12.3	23.8	17.9	19.4

Seven variables were used to measure quality at the organizational level of maintenance. In Table IV, the mean and standard deviation for each variable is shown for each of the three classes of ships.

The definition of the first three variables is self-evident. Explanations of the last four variables follow. "Delay" jobs are corrective maintenance jobs which have been deferred because of ship's operations, lack of repair parts, or the requirement for outside assistance. This type of job requires both initiating and closing cards. Absence of a closing card was defined as incomplete documentation. A "parts" job is one wherein the need for a part is indicated. Documentation of such a job requires at least a maintenance action card and one part card. Absence of a part card indicates incomplete documentation. Variable x_6 is the percentage of parts cards for which the associated maintenance action cards could not be found. A "work request" job is one requiring outside assistance. Complete documentation of this type of job calls for the submission of an initiating deferral card indicating the need for outside assistance. Incomplete documentation was indicated when this deferral card was absent.

In Table IV, the mean values of each of these qualitative variables are shown for the three classes A, B, and C as defined and used in the regression analysis previously described.* As can be seen, the mean values change in the expected direction. Of interest from the point of view of the assumptions implicit in the use of the discriminant analysis, the corresponding standard deviations of the selected measurements are approximately the same size across the three groups.

Given the set of seven quality measures just described, the discriminant analysis was used to assign each ship to exactly one of three classes, a, b, and c. Since the program used for this analysis considered only two classes at a time and three classes of ships were recognized, the total sample of ships was first divided into "potential a" ships and "potential c" ships. Using the discriminant analysis a second time, ships in the former groups were then assigned to class a or class b; likewise, ships in the latter group were assigned to class b or class c. The results of this process are shown in Table V where ships are classified on the basis of card volume and the application of the discriminant analysis.

TABLE V

Assignment: Card volume	Assignment: Discriminant function			
	a	b	c	Total
A	14	3	1	18
B	6	19	5	30
C	0	4	10	14
Total	20	26	16	62

The first observation to be made on the basis of Table V is the high concentration of ships along the main diagonal. The percentage of ships classified correctly was 78 among A ships, 63 among B ships, and 71 among C ships. The second observation to be made is that among the ships initially categorized as being C ships in terms of card volume, none was assigned to group a by the discriminant analysis. This is especially important since in the selection of a sample of better reporting ships there is a high cost associated with including poorer reporting ships. On the other hand, only one A ship was assigned to the c group. Based on these results, it was concluded that there were marked differences among ships in terms of both quantity and quality of reporting and that discriminate analysis can be an extremely useful procedure for differentiating between better and poorer reporters.

Granting that substantially the same assignment of ships into classes is obtained by using a discriminant function based on measures of quality and by using empirical evidences based on card count, is there any advantage to employing the former and, if so, how can costs be minimized without sacrificing this advantage?

With respect to the first part of this question, it may be concluded that there is an advantage in applying discriminant analysis in the manner indicated. It permits the treatment of an important aspect of adequate reporting, namely, the quality of reporting.† For example, among the sample ships originally rated as A ships on the basis of card volume alone, one ship was consistently excluded from

*The classes A, B, and C are similarly defined in Tables V and VI.

†For this paper, the clustering procedure was applied only to the question of volume of reporting. Clustering procedures, however, can be applied to variables representing both quantity and quality of reporting.

class a using discriminant functions. This ship was the one with the lowest card volume among all A ships. From the results of this study, it would appear that by requiring reporting units to qualify on the bases of both quantity and quality of reporting, marginal units can be eliminated from the category designated as better reporters.

A partial answer to the second part of the question posed above, i.e., the choice of measurements to minimize the cost of obtaining information, can be obtained by comparing the degree to which ships can be differentiated using different combinations of quality measurements. If it is required that ships be judged good reporters with respect to both quantity and quality of reporting, attention can be restricted, say, to the cells (A, a), (A, b), (B, a), and (B, b) in Table V. If instead of using all the quality measurements as in Table VI(c), only variables x_1 and x_6 are used, the assignment of ships is as shown in Table VI(a). Adding the additional two variables x_2 and x_3, for example, leads to the assignment shown in Table VI(b). As may be seen from these matrices, as more variables are employed, the concentration of ships represented on the main diagonal increases.* It appears that as few as four measurements of quality would suffice if ships were required to meet both quantity and quality criteria for good reporting.

TABLE VI. *Discriminant Functions*

		a	b	Total
Variables: x_1, x_6				
(a)	A	15	3	18
	B	14	2	16
	Total	29	5	34
Variables: x_1, x_2, x_3, x_6				
(b)	A	13	5	18
	B	7	17	24
	Total	20	22	42
Variables: x_1, x_2, . . . , x_7				
(c)	A	14	3	17
	B	6	19	25
	Total	20	22	42

CONCLUSION

In this paper we have examined some statistical techniques for auditing large-scale management information systems. While these techniques are well known, they have not been applied previously to the problem of auditing information. What is likely to be the usual approach is to examine a large

*The change in the total number of ships among matrices is due to the assignment of fewer A and B ships to class c as more quality variables are added.

number of individual data elements for format or similar errors. However, this task, although necessary, is not sufficient where little or nothing is known about the underlying characteristics of the process to be monitored. The approach used in this paper was to address the prior problem of identifying good vs poor reporters as a first step approximation to identifying good vs poor data.

The techniques of regression analysis, cluster analysis, and discriminant analysis were applied to a particular set of data which is illustrative of many of the problems met in the important initial phase of auditing large-scale information systems. For the system examined, i.e., the Navy Maintenance Data Collection System, large differences in level of reporting of maintenance activity among ships were observed. These differences in card volume were used as a first step in identifying "good" vs "bad" reporters.

The use of regression analysis as an auditing technique was dictated by the need to determine whether the observed variations in card volume could be explained in terms of variables that had very little to do with quality of data reporting, e.g., size of ship, age of ship, hours underway, etc., but which may have been important in explaining the volume of reporting. The findings suggested that operational characteristics explained only a portion of the variation in card volumes.

Cluster analysis, although not described in detail here, was employed as an independent check to determine if operational and maintenance activity variables were related to card count without imposing an *a priori* structure on the data. This check was deemed useful since, in part, the regression findings depended on the initial classification of ships into groups.

The discriminant analysis was used for several reasons, each of which suggesting that this was an appropriate technique given the context of the study. First, like cluster analysis, discriminant analysis groups observations rather than requiring the investigator to do this on the basis of his preconceived notions. Second, it permits evaluation of how well a set of variables discriminates among elements with known characteristics. As noted, the "known" characteristics pertaining to quality of reporting were represented by a proxy measure based on card count. For our purpose this was satisfactory, since we were interested in the extent to which card count and other measures of quality of reporting were correlated. Third, it permits the computation of a scalar index from a vector of measurements each representing a different aspect of reporting quality. Finally, we wished to illustrate the usefulness of discriminant analysis in identifying outlier observations. We conclude that discriminant analysis is a useful technique for ranking activities in terms of quality of reporting. We found that using a small number of qualitative measurements, the results were similar to those obtained by using card count alone as an indicator of quality.

For the particular system studied, and perhaps for other systems as well, the finding of covariance between volume and quality (other than that indicated by volume) of reporting was felt to be particularly important from an operational point of view. It should be noted that had the outcome been otherwise, the card volume measure would have been rejected and only the discriminant analysis would have been offered as a large-scale auditing procedure.

ACKNOWLEDGMENTS

The authors are indebted to Professor Rosedith Sitgreaves for invaluable technical advice. Also note should be made of the excellent programming assistance provided by Messrs. Richard W. Carr, Steven A. Frank, and William G. Underhill.

REFERENCES

[1] Fisher, R. A., "The Use of Multiple Measurements in Taxonomics Problems," Annals of Eugenics 7, 179–188 (1936).

[2] King, Benjamin, "Step-wise Clustering Procedures," J. Am. Statistical Assoc. 62, 86–101 (1967).

[3] Smith, Cedric A. B., "Some Examples of Discrimination," Annals of Eugenics 13, 272–282 (1947).

AN ECONOMIC MODEL FOR PLANNING STRATEGIC MOBILITY POSTURE*

G. Bennington

North Carolina State University†

and

L. Gsellman and S. Lubore

The MITRE Corporation

ABSTRACT

This paper describes a linear programming model used to perform cost-effectiveness analysis to determine the requirements for military strategic mobility resources. Expenditures for three classes of economic alternatives are considered: purchase of transportation resources; prepositioning of material; and augmentation of port facilities. The mix of economic alternatives that maximizes the effectiveness of the resultant transportation system is determined for a given investment level. By repeating the analysis for several levels of investment, a cost-effectiveness tradeoff curve can be developed.

INTRODUCTION

This paper describes a linear programming model which has been used by the Organization of the Joint Chiefs of Staff to determine the requirements for strategic mobility resources. The model can be used to examine a set of economic alternatives and to determine the mix of alternatives which provide maximum effectiveness for each of several system costs, i.e., a cost-effectiveness evaluation.

Several papers address the strategic mobility planning problem. Hitch and McKean describe the strategic mobility problem and present a procedure for choosing an intercontinental military air transport fleet [4]. Mihram develops a simulation approach for selecting such an air fleet [6]. Gsellman and Sandler describe an approach for the investigation of the operations planning phase of strategic mobility problems [3].

Lynn considers both aircraft and ships in his approach to the economic analysis of strategic mobility resources [5]. Lynn presents, in broad terms, a linear programming model for performing the analysis. A detailed description of a linear programming model based on the framework specified by Lynn is presented in papers by Fitzpatrick, Bracken, O'Brien, Wentling, and Whiton [2].and by Whiton [7]. Basically, the formulation presented by the above authors provides the capability of determining the least-cost mix of mobility resources to satisfy time-phased military shipping requirements.

The model described in this paper includes a number of additional economic alternatives which have not previously been incorporated into the models used by the Organization of the Joint Chiefs of Staff in their analyses of strategic mobility systems.

---

*This paper was presented at the 40th National ORSA Meeting, October 27–29, 1971. The work was sponsored by the Defense Communications Agency under contract F 19628–68–C–0365.

†This work was performed while the author was at the MITRE Corporation.

MODEL DESCRIPTION

Strategic mobility is the capability to perform the movement of military requirements—men, equipment, and supplies—over channels from points of origin to points of end use in a timely fashion and in accordance with a desired sequence of arrivals. The model described in this paper is a resource allocation model structured as a linear program which addresses the intertheater portion of this problem, i.e., the movement by transportation resources of military requirements between ports of embarkation (POE's) and ports of debarkation (POD's). The movement of requirements from the points of origin to the POE's and from the POD's to the points of end use is not considered.

Time is considered in the model, by dividing the total time of a deployment into time periods, each having a length of two or more days. To ensure feasible solutions to the linear program, a dummy time period and a dummy vehicle are included. Any requirements which cannot be allocated by the model to vehicles during feasible time periods can always be allocated in the dummy time period to the dummy vehicle. Allocations in the dummy time period are the requirements which cannot be moved, i.e., requirements that are in "shortfall".

The model is composed of two types of decision variables: operational variables and economic variables. The operational variables are those variables which represent the movement of a requirement. Operational variables are

$$X_{i,j,k,m}$$

and represent the amount of requirement i moved in time period j by vehicle type k on channel m.

The economic variables fall into three classes and describe actions taken prior to initiation of a deployment. These variables are:

(a) transportation resource variables Z_k, which represent the numbers of additional vehicles of each type k purchased

(b) pre-positioning variables $P_{i,m}$, which represent the amounts of each requirement i prepositioned to each channel

(c) port facility variables Y_n, which represent the amounts of additional capacity of each port n purchased

The objective function used in the model represents the system effectiveness and includes only the operational variables. The lower the value of the objective function, the larger the effectiveness of the system. As a result, maximum effectiveness solutions are obtained by minimizing the objective function. The general form of the objective function is

$$\sum_{i \in I} \sum_{m \in M_i^3} \sum_{k \in K_m} \sum_{j \in J^1_{i,k,m}} (W_{i,j,k,m})(X_{i,j,k,m}), \tag{1}$$

where

$W_{i,j,k,m}$ is the weight associated with the operational variable $X_{i,j,k,m}$.

I is the set of all requirements.

M_i^3 is the set of all feasible channels for the movement of requirement i.

K_m is the set of all vehicle types that can operate on channel m.

$J^1_{i,k,m}$ is the set of time periods in which a vehicle of type k can leave the POE of channel m and deliver requirement i within its specified delivery time span to the POD of channel m.

The weight assigned to each operational variable $X_{i,j,k,m}$ is a function of the requirement being moved, the delivery date, the vehicle being used, and the channel selected.

Several weighting functions have been used; however, the one used most frequently is the one that results in shortfall as a measure of effectiveness. In this case, $W = 0$ for all time periods other than those associated with the dummy time period. For the dummy time period, $W = 10{,}000$. By use of this weighting function, the model objective is the minimization of shortfall.

Other weighting functions have been used which attempt to minimize delivery time or minimize deviations about the desired delivery dates. Bennington and Lubore describe several weighting functions which reflect preferences for any or all of the factors previously mentioned [1].

The constraints placed upon the variables used in the model are divided into operational constraints and program planning constraints. The operational constraints reflect physical limitations placed upon the utilization of system resources. The operational constraints are divided into:

(a) requirement constraints
(b) vehicle constraints
(c) port constraints
(d) an operating cost constraint.

The program planning constraints are used in conjunction with the operational constraints to provide a means by which the model can make tradeoffs between various economic alternatives. Program planning constraints fall into three groups:

(a) an investment constraint
(b) pre-positioning constraints
(c) peacetime economic value constraints.

The investment constraint limits the system cost to a specified amount. By varying the right-hand side of this constraint, a series of solutions are obtained which describes the system and its effectiveness for each of several costs, i.e., a cost-effectiveness tradeoff curve.

The peacetime economic value constraints provide the model with the ability to consider the impact of peacetime uses of government-owned resources on the net cost of the transportation system.

Requirement Constraints

The model contains two types of constraints upon the movement of transportation requirements. One set of constraints (primary constraints) ensures that all requirements are either moved or prepositioned. A second set of constraints (prepositioning constraints) ensures that all prepositioned requirements are moved. The requirement constraints can also provide for accompanying cargo. Accompanying cargo is used when a vehicle can carry one type of cargo without necessitating reduction of the amount of other types of cargo which it can carry. For example, a C–5A aircraft can carry up to a given number of passengers without restricting the cargo tonnage which it can carry.

The primary requirement constraints state that the total amount of each requirement is equal to the amount shipped directly over a feasible channel plus the amount prepositioned on a feasible prepositioning channel.

If the accompanying cargo option is not employed, primary requirement constraints are:

$$\sum_{m \in M_i^1} \sum_{k \in K_m} \sum_{j \in J_{i,k,m}^1} X_{i,j,k,m} + \sum_{m \in M_i^2} P_{i,m} = R_i, \qquad i \in I, \tag{2}$$

where

$P_{i,m}$ is the amount of requirement i prepositioned to channel m.

M_i^1 is the set of feasible channels over which requirement i can be moved directly.

M_i^2 is the set of channels to which requirement i can be prepositioned. The set is empty if requirement i cannot be prepositioned.

R_i is the amount of requirement i to be moved.

The prepositioning variables $P_{i,m}$ in Eq. (2) insure that all prepositioned material is allocated to the prepositioning sites. The mathematical equations which insure that the prepositioned material is moved into the theater of destination are*:

$$\sum_{k \epsilon K_m} \sum_{j \epsilon J^1_{i,k,m}} X_{i,j,k,m} - P_{i,m} = 0, \qquad i \epsilon I,\ m \epsilon M_i^2. \tag{3}$$

The existence of accompanying cargo requires a modification to the requirement constraints Eqs. (2) and (3).

The amount of each requirement moved by normal loading of vehicles or as accompanying cargo, and the amount prepositioned must be greater than or equal to the amount desired. The resulting constraint is:

$$\sum_{m \epsilon M_i^1} \sum_{k \epsilon K_m} \sum_{j \epsilon J^1_{i,k,m}} \sum_{i' \epsilon I} (D_{i,i',j,k,m})\ (X_{i,j,k,m}) + \sum_{m \epsilon M_{i'}^2} P_{i',m} \geq R_{i'}, \qquad i' \epsilon I, \tag{4}$$

where

$D_{i,i',j,k,m}$ is the amount of requirement i' accompanying a unit of requirement i on vehicle type k if shipped in time period j on channel m. (Note that $D_{i,i,j,k,m} = 1$).

The mathematical representation of the prepositioning constraints with accompanying cargo is:

$$\sum_{k \epsilon K_m} \sum_{j \epsilon J^1_{i,k,m}} \sum_{i' \epsilon I} (D_{i,i',j,k,m})\ (X_{i,j,k,m}) - P_{i',m} = 0, \qquad i' \epsilon I,\ m \epsilon M_{i'}^2. \tag{5}$$

Vehicle Constraints

The vehicle constraints insure that the numbers of each vehicle type used in each time period do not exceed the numbers available, and consider both existing resources and purchased resources. The constraints also insure that any requirements that are moved use lift resource capacity, and that this capacity is subtracted from the total resource capacity available during the time resources are being used either to deliver requirements or return to the POE's. After a round trip, a vehicle is assumed to be available for reuse at any POE. The times for vehicles to return to POE's different from the ones they left are not considered. The mathematical representation of the vehicle constraints is:

*An alternative formulation is to include the first terms of Eqs. (3) in Eqs. (2) and to drop the $P_{i,m}$ variables. In this formulation, constraints Eqs. (3) do not exist; thereby reducing the size of the model. The formulation presented is used for expository purposes.

$$(6) \qquad \sum_{i\epsilon I} \sum_{m\epsilon M_i^3} \sum_{j'\epsilon J_{j,k,m}^2} \frac{X_{i,j',k,m}}{A_{ij',k,m}} - Z_k = V_{j,k}, \qquad j\epsilon J,\ k\epsilon K,$$

where

$J_{j,k,m}^2$ is the set of time periods that a vehicle of type k can leave on channel m and is still in transit during time period j. (Note: the set $J_{j,k,m}^2$ always includes time period j.)

$A_{i,j,k,m}$ is the carrying capability of one vehicle of type k to transport requirement i on channel m during time period j.

Z_k is the number of additional vehicles of type k which are purchased.

$V_{j,k}$ is the number of vehicles of type k available during time period j.

K is the set of all vehicle types.

The calculation of $A_{i,j,k,m}$ falls into one of two cases:

CASE I: $J_{j,k,m}^2 = j$.

In this case, a vehicle of type k has a round trip transit time less than the duration of time period j when traveling on channel m. The factor $A_{i,j,k,m}$ represents the productivity of the vehicle during time period j considering the vehicle's average utilization rate and reflects the possibility of a vehicle making more than one trip or "sortie" during the time period.

CASE II: $J_{i,k,m}^2$ contains more than one time period.

In this case, a vehicle of type k requires more than one time period to make a round trip over channel m. In this case, $A_{i,j,k,m}$ represents the capacity of a vehicle of type k to transport requirement i on channel m in time period j.

Port Constraints

The port constraints insure that the capacity, in terms of vehicles per time period, of any port (POE, POD, or en route base) in any time period is not exceeded. The constraints apply to both air and sea ports. POD's and en route bases are constrained on arrivals of vehicles; POE's are constrained on departures of vehicles.

The mathematical expression for the port constraints is:

$$(7) \qquad \sum_{i\epsilon I} \sum_{m\epsilon M_{i,n}^1} \sum_{k\epsilon K_{m,n}} \sum_{j'\epsilon J_{j,k,m,n}} \frac{X_{i,j',k,m}}{B_{i,j',k,m}} - (L_j)\,(Y_n) \leq F_{j,n}, \quad n\epsilon N,\ j\epsilon J,$$

where

$J_{j,k,m,n}$ is the set of time periods in which a vehicle of type k operating on channel m must leave to arrive at port n during time period j.

N is the set of all constrained ports.

$M_{i,n}^1$ is the set of feasible channels for requirement i which contain port n.

$K_{m,n}$ is the set of vehicle types that operate on channel m and pass through port n.

$B_{i,j',k,m}$ is the capacity of one vehicle of type k arriving at port n in time period j carrying requirement i leaving the POE of channel m in time period j'.

$F_{j,n}$ is the capacity of port n in vehicles per time period in time period j.
Y_n is the amount of additional capacity per day of port n which is purchased.
L_j is the duration in days of time period j.

Equations (7) limit the number of vehicles passing through each port n in each time period j to the sum of the original port capacity $F_{j,n}$ and the purchased port capacity $(Y_n)(L_j)$.

Operating Cost Constraint

The operating cost constraint restricts the dollar operating cost of performing the deployment to less than or equal to a given amount. The operating cost constraint is given by

$$\sum_{i \in I} \sum_{m \in M_i^3} \sum_{k \in K_m} \sum_{j \in J_{i,k,m}^1} (C_{i,j,k,m})(X_{i,j,k,m}) \leq G, \tag{8}$$

where

$C_{i,j,k,m}$ is the cost in \$/unit of moving requirement i in time period j with a vehicle of type k on channel m.
G is the operating budget in \$ available to support the movement.

Investment Constraint

The investment constraint limits the dollars spent on the system for new resources to a given amount. In allocating the investment dollars, the model chooses between expenditures for prepositioning, additional lift resources, or augmentation of port capacity. The constraint also reflects any revenue which might be derived from the peacetime use (peacetime economic value) of the lift resources purchased. The equation for the investment constraint is

$$\sum_{k \in K} (C_k^1)(Z_k) + \sum_{n \in N} (C_n^2)(Y_n) + \sum_{i \in I} \sum_{m \in M_i^2} (C_{i,m}^3)(P_{i,m}) - \sum_{s \in S} T_s \leq Q, \tag{9}$$

where

C_k^1 is the cost of procuring additional vehicles of type k in \$/vehicle.
C_n^2 is the cost of procuring additional capacity for port n in \$/vehicle arrival per day (\$/vehicle departure per day at POE's).
$C_{i,m}^3$ is the cost in \$/unit of prepositioning equipment i to channel m.
T_s is the value of using the purchased vehicles to deliver peacetime requirements. This value is derived from segment s of the PEV curve (Figure 1).
S is the set of segments of the PEV curve.
Q is the maximum amount of the investment in \$ available.

A more detailed discussion of the peacetime economic value (PEV) is presented below. For now, all that is necessary is to note that PEV is a credit against the cost of the system.

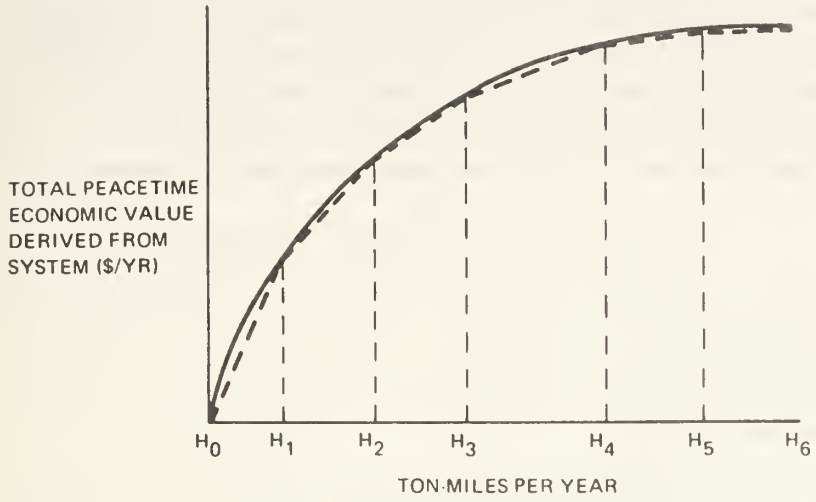


Figure 1. Linear Approximation of Total Peacetime Economic Value (PEV) Curve

Common Prepositioning Constraints

Common prepositioning arises when material can be prepositioned and used to satisfy the requirements of more than one nonsimultaneous contingency. Under these conditions it may be economically advantageous to preposition material for a contingency and permit the same material to be used also to support other contingencies without incurring any additional expenditure.

Common prepositioning is handled by setting the prepositioning cost of all but the largest common prepositioning requirements to zero and adding "limit equations."

Let I_g be the gth set of requirements for the nonsimultaneous contingencies which are under consideration for common prepositioning. These requirements must be prepositionable and substitutable, and must have a common prepositioning site, i.e., a port such that requirements in the set I_g can be prepositioned to channels originating at the common site.

Let the set $M_{i,}{}^{2}{}_{n}$ include those channels for requirements i which originate at port n. Let $I_{g,\,n}$ be the subset of a set of requirements I_g which can be prepositioned to common site n. i.e.,

$$I_{g,\,n} = \left\{ i \epsilon I_g \; M_{i,}{}^{2}{}_{n} \neq \Phi \right\},$$

where

Φ = the null set. (Note: the set $I_{g,\,n}$ must have at least two elements for common prepositioning to be considered.)

Let i' be the requirement with the largest amount of substitutable material which can be prepositioned to the common site, i.e.,

$$R_{i'} = \underset{i \epsilon I_{g,n}}{\text{maximum}} \; [R_i].$$

For each common site, a cost coefficient for each variable $P_{i',\,m}$ corresponding to the preposition-

ing of requirement i' to each channel in the set $M^2_{i',n}$ is included in the investment constraint Equation (9). All other requirements in the set $I_{g,n}$ are prepositioned without a cost penalty. The approach is valid if the amount of each free common requirement is less than or equal to the total amount of requirement i' prepositioned to site n. To insure that this condition holds, a set of "limit" constraints are used. These constraints for the set of common requirements I_g and common site n are:

$$\sum_{m\epsilon M^2_{i',n}} P_{i',m} - \sum_{m\epsilon M^2_{i,n}} P_{i,m} \geq 0, \qquad (i \neq i')\epsilon I_{g,n}. \tag{10}$$

Sea Prepositioning Constraints

Sea prepositioning uses ships as prepositioning sites. Only vehicles designated as sea prepositioning vehicles may be used in this mode.

If a sea prepositioning activity takes place, the vehicle involved is considered unavailable for other service until the vehicle delivers its prepositioned cargo. This differs from the normal procedure in which a vehicle is available to the system until it is loaded.

Irrespective of whether common sea prepositioning is employed, sea prepositioning changes only the constraints on the vehicles which may be used for sea prepositioning. The set $J_{j,k,m}$ in Equation (6) is replaced by a new set $J^3_{j,k,m}$ which is the set of all time periods from the first time period until the first time period until the vehicle delivers the prepositioned cargo and returns to its original site. Modifying Equation (6) accordingly results in:

$$\sum_{i\epsilon I}\left(\sum_{m\epsilon(M^3_i - M^4_i)}\sum_{j'\epsilon J^2_{j,k,m}} \frac{X_{i,j',k,m}}{A_{i,j',k,m}} + \sum_{m\epsilon M^4_i}\sum_{j'\epsilon J^3_{j,k,m}} \frac{X_{i,j',k,m}}{A_{i,j',k,m}}\right) - Z_k \leq V_{j,k}, \qquad j\epsilon J, k\epsilon \hat{K}. \tag{11}$$

where

M^4_i is the set of sea prepositioning channels to which requirement i can be prepositioned.

$J^3_{j,k,m}$ is the set of all time periods from the first time period until vehicle type k completes the delivery on sea prepositioning channel m and returns to its original site.

$\hat{K}$ is the only set of vehicle types used for sea prepositioning.

Peacetime Economic Value Constraints

The model is capable of including the peacetime economic value generated when government-owned vehicles are used to deliver military cargo during peacetime.

In general, the more cargo moved or ton-miles produced, the smaller the savings observed on each additional ton-mile produced. A typical return function is presented in Figure 1. The curve represents the total dollars/year saved as a function of the total ton-miles per year produced as a result of using government owned vehicles. The PEV is the difference of the return and the operating expense.

To represent the PEV functions, the curves are approximated by linear segments. Associated with each segment is an end point and a slope, i.e., dollars of PEV per ton-mile. The PEV curve is always concave downward, i.e., the slopes of the segments are monotonically decreasing.

The model utilizes three types of constraints to include PEV. The first, a set of transfer equations, calculates the number of dollars of PEV derived from using available vehicles and provides the basis

for transferring this PEV to the investment constraint. The PEV transferred to the investment constraint Equation (9) is used as a credit against other investments.

The second set of constraints limits the number of vehicles used in the calculations of the transfer equation to the number available in the system.

The third set of constraints limits the number of ton-miles produced by all government owned vehicles at a given PEV rate.

The interaction of the three PEV constraints insures that vehicles are credited with any economic value associated with the peacetime utilization of the vehicles subject to limitations in the number of vehicles available and the amount and value of peacetime work to be performed.

Transfer of the PEV derived from the peacetime use of lift resources is expressed mathematically as

$$\sum_{k \epsilon \tilde{K}} (C^4_{k,s})(U_{k,s}) - T_s = 0, \qquad s \epsilon S, \tag{12}$$

where

$\tilde{K}$ is the set of vehicles which contributes to the PEV.

$C^4_{k,s}$ is the amount of dollars of PEV derived from operating for 1 year one vehicle of type k on segment s of the PEV curve and is the difference between the revenue obtained by the peacetime use of the vehicle and the cost of operating the vehicle to obtain that revenue.

S is the set of segments of the PEV curve.

$U_{k,s}$ is the number of vehicles of type k used for delivery of peacetime cargoes associated with segment s of the PEV curve.

In order to insure that resources used to provide PEV do not exceed the number available, a PEV vehicle constraint is employed for each vehicle type under consideration for purchase. This constraint guarantees that no more vehicles will be used in producing PEV than are actually available (V_k) or can be purchased (Z_k). The constraint is

$$\sum_{s \epsilon S} U_{k,s} - Z_k \leq V_k, \qquad k \epsilon \tilde{K}. \tag{13}$$

Because the PEV/ton-mile changes for each segment of the PEV curve, a PEV ton-mile constraint is necessary. This constraint restricts the number of ton-miles produced at a given PEV rate to the number of ton-miles available at that rate as indicated by the PEV curve. Without this constraint all PEV would be generated at the highest rate, i.e., the rate associated with the first segment of the PEV curve. The mathematical representation of this constraint is

$$\sum_{k \epsilon \tilde{K}} (E_k)(U_{k,s}) \leq (H_s - H_{s-1}), \qquad s \epsilon S, \tag{14}$$

where

E_k is the annual ton-mile capability of vehicle type k.

H_s is the maximum annual ton-miles flown corresponding to segment s of the PEV curve (note that $H_0 = 0$).

IMPLEMENTATION

The model described in the preceding paragraphs has been used to support several planning studies. Because use of the model to support these studies has resulted in large linear programming problems, i.e., up to 1100 constraints and 40,000 variables, a preprocessor was designed to generate the matrix. In addition, a post-processor was designed to aid in the analysis of the solutions obtained from the model. The preprocessor and the post-processor were written in FORTRAN and designed to operate in conjunction with IBM's mathematical programming system. Typical problems of 400–500 constraints and 20,000 variables requires 40–90 minutes of IBM 360/50 computer processing. These times include matrix generation, linear program solution, and preparation of reports and post-processing.

REFERENCES

[1] Bennington, G. and S. Lubore, "Resource Allocation For Transportation," Nav. Res. Log. Quart. **17,** 471–484 (Dec. 1970).

[2] Fitzpatrick, G. R., J. Bracken, M. J. O'Brien, L. G. Wentling, and J. C. Whiton, "Programming the Procurement of Airlift and Sealift Forces: A Linear Programming Model for Analysis of the Least-Cost Mix of Strategic Deployment Systems," Nav. Res. Log. Quart. **14,** 241–255 (June 1967).

[3] Gsellman, L. R., and M. M. Sandler, "Economic Analysis of the Operations Planning Phase of Strategic Mobility Problems," in *Business Logistics: Policies and Decisions*, D. McConaughy and C. J. Clawson, editors (the University of Southern California, Los Angeles, California, 1968), pp. 157–177.

[4] Hitch, C., and R. McKean, *The Economics of Defense in the Nuclear Age* (Harvard University Press, Cambridge, Massachusetts, 1960), pp. 133–158.

[5] Lynn, L. E., Jr., "Economic Models in the Analysis of Military Strategic Mobility and Transportation Requirements," presented at the Conference on the Analysis of Systems of Military Transportation, Oxford, England, July 1966.

[6] Mihram, G. A., "A Cost-Effectiveness Study for Strategic Airlift," Transportation Science **4,** 79–96 (Feb. 1970).

[7] Whiton, J. C., "Some Constraints on Shipping in Linear Programming Models," Nav. Res. Log. Quart. **14,** 257–260 (June 1967).

A MIXED OPTIMIZATION TECHNIQUE FOR THE GENERALIZED MACHINE REPLACEMENT PROBLEM*

S. P. Sethi

and

T. E. Morton

Graduate School of Industrial Administration
Carnegie-Mellon University

ABSTRACT

A mixed optimization technique for optimal machine replacement is presented which allows much more flexibility than previous models. Optimal purchase, maintenance and sale of a given machine between any two given points in time is treated as a subproblem, which one may choose to solve via control theory, dynamic programming, or practical engineering considerations. (A control theory formulation is used in the paper as an illustration.) These subproblem solutions are then incorporated into a Wagner-Whitin formulation for solution of the full problem. The technique is particularly useful for problems with such asymmetries as an existing initial machine or uneven technological change. A simple numerical example is solved in the Appendix.

I. INTRODUCTION

The first major work dealing with machine replacement problems appeared in 1949 as a MAPI study by Terborgh [7]. For the most part, this study was confined to those problems where the optimization was carried out only with respect to the replacement lives of the machines under consideration. In 1955, Boiteux (see Masse [3]) extended the single machine replacement problem to include the optimal timing of a partial replacement of the machine before its actual retirement. After that, not much was done, except for some simple dynamic programming formulations [9], until Naslund [4] and Thompson [8] solved a generalized version of the Boiteux problem by using Pontryagin's maximum principle. They considered optimal preventive maintenance over the entire period instead of a single optimal partial replacement before the machine is retired. Sethi [5] extended Thompson's model [8] by considering optimization over constant infinite chains of replacements under a changing technological environment. While useful in certain special situations,† the assumption of constant infinite chains (i.e. chains with identical links) is quite restrictive, and especially so under a changing technological environment.

In this paper we deal with a finite horizon replacement model which can be put into perspective with the past literature as follows.

First, the model relaxes the assumption of identical links in [5] at the cost of making the replacement chain finite; second, it generalizes Terborgh's problem (where machines in the chain are allowed to

*This report was prepared as part of the activities of the Management Sciences Research Group, Carnegie-Mellon University, under Contract N0014-67-A-0314-007 NR 047-048 with the U.S. Office of Naval Research. Reproduction in whole or in part is permitted for any purpose of the U.S. Government.

†For example, the replacement problem of a firm maintaining a large fleet of fork lift trucks scattered at several of its locations.

Also see Masse [3, page 74] . . . "the assumption of one constant chain is admissible only in applications where the future is far enough removed to allow discounting factors to palliate their effort on the present to a sufficient extent."

have unequal lives [3, 8]) to include simultaneous determination and optimal preventive maintenance policy over the life of each machine in the chain.*

Further, the model considered in this paper can be solved at any time to decide whether to replace the existing machine at that time. In other words, the model does not require that we start the chain with a new machine. This also implies that if the parameters of the problem change at some future time, the problem can be re-solved.

A basic insight of the solution technique proposed here is that the optimal maintenance policy for a given machine depends only on its purchase date and the purchase date of its successor. Each such individual problem may be conceptually solved in a number of possible ways. In this paper, the discrete maximum principle has been utilized. However, for certain types of smooth functions, the continuous version might be computationally simpler, or maintenance policy might be dictated by engineering considerations, or even by "rules of thumb" for problems not justifying more extensive calculations. Thus the discussion in the next section should be considered as illustrative of one major approach to the single-machine subanalysis, rather than as definitive.

The problem of determining an optimal chain of such machines, on the other hand, is strongly reminiscent of the Wagner-Whitin dynamic lot size model [10]. The expense incurred in buying a machine corresponds to the setup cost in initiating production run: revenue, maintenance, and salvage costs correspond to production and holding costs.

Use of the Wagner-Whitin algorithm requires: i) a finite horizon, ii) a finite number of decision points, and iii) concave subproblem costs. The first assumption gives no difficulty, since a 50- or 100-year horizon is adequate for all practical purposes. Choice of a finite number of time points to allow buying a machine is conceptually easy also; certainly 1 year increments would be adequate for many types of problems. However, since the number of subproblems to be solved goes up as the square of the number of such decision points, in practice a judicious selection would be advantageous. A crude analysis might first suggest the rough time at which the current machine might be expected to be retired; extra decision points could be inserted during this period. Wider intervals could be used farther in the future; expected introduction points for new technologies should be included. (We have here followed the simpler procedure of using equal intervals.) Finally, we have assumed away the problem of concavity by simply requiring that only one machine be used at a time.† This restriction is really critical for the solution procedure; fortunately it should often be closely approximated in practice.

II. CONTROL THEORY FORMULATION OF THE SIMPLE SINGLE MACHINE SUBPROBLEM

Consider buying a machine at the beginning of period n and salvaging it at the end of period k. Let V_{nk} denote the present value of all net earnings associated with the machine.‡ To calculate V_{nk} we need the following.

State Variables at t, $n \leq t \leq k$

$S(t)$; Salvage value of the machine at the end of period t

$$R(t) = P(t) - E(t),$$

*Charles Tapiero of Columbia University is also working on a similar problem.

†That is to say, we don't care if each subproblem is concave or not, since we have required the associated network to be arborescent, because only one machine may be used at a time.

‡It is also easy to modify the analysis to allow periods with no machine if desired.

where

$P(t)=$ production (in dollar value) during period t, and
$E(t)=$ necessary expense of the ordinary maintenance (in dollars) during period t.

Control Variable

$m(t)$, Preventive maintenance (in dollars) carried out at the end of period t, $n \leq t \leq k-1$; (clearly $m(k)=0$).

It is required that $0 \leq m(t)$ 4 $U_n(t) \forall t \, e\,[n,k]$.

Exogenous Functions

C_n cost of purchasing machine at n.

Assuming r as the discounting rate with $\rho=(1+r)$, we can now calculate V_{nk} in terms of the functions defined above.

$$V_{nk}=\sum_{t=n}^{k} R(t)\,(1+r)^{-t}-\sum^{k} m(t)\,(1+r)^{-t}-C_n\,(1+r)^{-n}+S(k)\,(1+r)^{-k}. \tag{1}$$

We must also have functions which will provide us with the way states change due to the age of the machine and the amount of preventive maintenance. Also, assuming that at time n, the only machines available are the ones up to date with respect to the technology prevailing at n, we can subscript these functions by n to reflect the effect of the machine's technology on its state at time t. Let $\psi_n\{m(t), t\}$ and $\phi_n\{m(t), t\}$ be these functions such that we can write the following state equations.

$$R(t+1)-R(t)=\psi_n[m(t), t], R(n)=R_n \tag{2}$$

and

$$S(t+1)-S(t)=\phi_n[m(t), t], S(n)=(1-\delta)C_n, \tag{3}$$

where δ is the fractional depreciation immediately after purchase of the machine.

To convert the problem into Mayer form, define:

$$A(t)=\sum_{i=n}^{t-1} (i)\,(1+r)^{-i} \tag{4}$$

and

$$M(t)=\sum_{i=n}^{t-1} m(i)\,(1+r)^{-i}. \tag{5}$$

Using Equations (4) and (5), we can rewrite (1) as

$$V_{nk}=R(k)\ (1+r)^{-k}+A(k)-m(k)(1+r)^{-k}-M(k)-C_n(1+r)^{-n}+S(k)(1+r)^{-k}. \tag{6}$$

The Mayer form of the problem is to maximize V_{nk} subject to the state Equations (2) and (3), and

$$A(t+1)-A(t)=R(t)(1+r)^{-t} \qquad A(n)=0 \tag{7}$$

and

$$M(t+1)-M(t)=m(t)(1+r)^{-t} \qquad M(n)=0. \tag{8}$$

Solution by Maximum Principle [1, 2]

We associate the influence function $Z_1(t+1)$, $Z_2(t+1)$, $Z_3(t+1)$, and $Z_4(t+1)$, respectively, with the state Equations (7), (8), (2), and (3), and form the variational Hamiltonian,

$$H=Z_1(t+1)\ R(t)(1+r)^{-t}+Z_2(t+1)\ m(t)(1+r)^{-t}+Z_3(t+1)\ \psi_n+Z_4(t+1)\ \varphi_n, \tag{9}$$

where the influence functions Z satisfy the following difference equations and terminal boundary conditions.

$$\Delta Z_1(t)=\frac{-\partial H}{\partial A}=0; \qquad Z_1(k)=1, \tag{10}$$

$$\Delta Z_2(t)=\frac{-\partial H}{\partial M}=0; \qquad Z_2(k)=-1, \tag{11}$$

$$\Delta Z_3(t)=\frac{-\partial H}{\partial R}=Z_1(t+1)\ (1+r)^{-t}; \qquad Z_3(k)=(1+r)^{-k}, \tag{12}$$

and

$$\Delta Z_4(t)=\frac{-\partial H}{\partial S}=0; \qquad Z_4(k)=(1+r)^{-k}, \tag{13}$$

where

$$\Delta Z(t) \triangleq Z(t+1)-Z(t).$$

The solution of these equations is

$$Z_1(t)=1, \tag{14}$$

$$Z_2(t)=-1, \tag{15}$$

$$Z_3(t)=\sum_{i=t}^{k}(1+r)^{-i}, \tag{16}$$

and

$$Z_4(t)=(1+r)^{-k}. \tag{17}$$

Z_1, Z_2, and Z_4 are constants for a fixed k.

To apply the maximum principle, we substitute (14–17) into the Hamiltonian (9), collect terms containing control m, rearrange and decompose H as:

$$H = H_1 + H_2\ (m), \tag{18}$$

where H_1 is that part of H which is independent of m, and $H_2(m)$ is

$$H_2(m) = -m(t)(1+r)^{-t} + \sum_{i=(t+1)}^{k} (1+r)^{-i}\psi_n + (1+r)^{-k}\varphi_n. \tag{19}$$

Here we apply the maximum principle to obtain the necessary condition for the optimal schedule of preventive maintenance expenditure in dollars. The condition of optimality is that H should be maximum along the optimal control path. If $m(t)$ were unconstrained, this condition is equivalent to setting the partial derivative of H, with respect to m, equal to zero.

$$\frac{\partial H}{\partial m} = \frac{\partial H_2(m)}{\partial m} = -(1+r)^{-t} + \frac{\partial \psi_n}{\partial m} \sum_{i=t+1}^{k} (1+r)^{-i} + \frac{\partial \varphi_n}{\partial m}(1+r)^{-k} = 0. \tag{20}$$

Equation (20) is a functional equation in m with the exception of particular case (which will be treated later in this section) when ψ_n and φ_n are linear in m. In general, (20) may or may not have a unique solution. For our case we will assume ψ_n and φ_n to be of the form such that we have unique solution for m. One such case occurs where ψ_n and φ_n are quadratic in m. In this case, (20) is linear in m and can be solved explicitly for a unique solution for m. Whenever a unique solution does exist, let this be

$$m(t) = U_{nk}(t). \tag{21}$$

The optimal control $m^*(t)$ (See [2]) is:

$$m^*(t) = \begin{cases} 0 & \text{if } U_{nk}(t) \geq U_n(t) \\ U_{nk}(t) & \text{if } 0 < U_{nk}(t) < U_n(t) \\ U_n(t) & \text{if } U_{nk}(t) \leq 0. \end{cases} \tag{22}$$

Special Case of Bang-Bang Control

This special case occurs when the problem is linear in the control. In this case, H can be maximized simply by having the control at its maximum when coefficient of m in H is positive and minimum when it is negative. This type of control is called bang-bang [2].

In our problem, if ψ_n and φ_n assume the form

$$\psi_n(m(t), t) = m(t)\psi_n(t), \tag{23}$$

$$\varphi_n(m(t), t) = m(t)\varphi_n(t), \tag{24}$$

then the coefficient of m in H, denoted by W, is

$$W = -(1+r)^{-t} + \psi_n \sum_{i=t+1}^{k} (1+r)^{-i} + \varphi_n(1+r)^{-k}, \tag{25}$$

and the optimal control m^* is given by

$$m^* = \begin{cases} 0 & \text{if } W < 0 \\ \text{Singular} & \text{if } W = 0 \\ U_n(t) & \text{if } W > 0. \end{cases} \tag{26}$$

Calculation of V^*_{nk}

Once we have obtained m^*, we can substitute it into (2) and (3) to obtain $R^*(t)$ and $S^*(t)$, which in turn can be used in (1) to obtain the optimal value of the objective function denoted by V^*_{nk}. This can be done for each n and $k \geq n$.

Starting with an Already Existing Machine

The above formulation can be easily adapted to the case where there is an existing machine at time 0. Let us call 0th period (i.e., $n = 0$) the starting period. Let T_{-1} denote the time at which machine was purchased.

$$T_{-1} \text{ - - - } \frac{0 \quad 1}{0 \quad 1 \quad 2}$$

Let $\hat{m}(t)$; $T_{-1} \leq t < 0$ be the actual preventive maintenance carried out (available from records). Then, we can rewrite (2) and (3) as follows:

$$R(t+1) - R(t) = \psi_{T_{-1}}(m(t), t),\ R(0) = R_{T_{-1}} + \sum_{t=T_{-1}}^{0} \psi_{T_{-1}}(\hat{m}(t), t). \tag{27}$$

$$S(t+1) - S(t) = \varphi_{T_{-1}}(m(t), t),\ S(0) = C_{T_{-1}} + \sum_{t=T_{-1}}^{0} \varphi_{T_{-1}}(\hat{m}(t), t). \tag{28}$$

Having redefined $R(t)$ and $S(t)$ and observing that $C_0 = 0$ on account of its being the sunk cost, we can rewrite (1) as

$$V_{0k} = \sum_{t=0}^{k} R(t)(1+r)^{-t} - \sum_{t=n}^{k} m(t)(1+r)^{-t} + S(k)(1+r)^{-k}. \tag{29}$$

It is easy to see from the above discussion that one can resolve the problem at any future time whenever a revised forecast of technology is available.

III. INCORPORATION INTO WAGNER-WHITIN FRAMEWORK FOR FULL PROBLEMS [10]

Let f_n denote the cost (discounted to period 0) of an optimal replacement and preventive maintenance policy for periods $n, n+1, \ldots, N$. Then,

$$f_n = \max_{k = n+1,\, n+2,\, \ldots\, N} [V_{nk}^* + f_k] \quad 0 \leq n \leq N, \text{ and} \tag{30}$$

$$f_{N+1} = 0. \tag{31}$$

The values of V_{nk}^* are available from the analysis in the previous section. The value of f_0 will give the required maximum.

We may remark that search in this model will be substantially reduced if we knew a priori that it was inoptimal to replace a machine before it is at least $L(n)$ periods old or after it is $0(n)$ periods old. In this case let $P(n) = \min [N, n + 0(n)]$.

$$f_n = \max_{k = n + L(n),\, n + L(n) + 1,\, \ldots\, P(n)} \ldots [V_{nk}^* + f_k]. \tag{32}$$

IV. CONCLUDING REMARKS

The mixed optimization technique presented here avoids many of the shortcomings of either pure dynamic programming or pure control theory formulations. Since the solution technique used to optimize a given machine represents a submodule of the overall method, the pure dynamic programming approach may be recognized as a special case. It should be advantageous, however, to be able to use methodology for the submodule most efficient for a given particular problem. Previous control theory formulations certainly do not seem adapted to the situation of an existing initial machine or other similar asymmetries.

This mixed technique can also be adapted to the case of probabilistic technological breakthroughs. Here the path of technological growth is assumed to be a tree with probabilities associated with its branches. The subproblems can be solved by using what we term a "conditional maximum principle." (This principle is discussed more completely in Reference [6].) However, the number of subproblems to be solved increases rapidly with the number of branches, thus putting computational limitations on the general usefulness of this extension.

Appendix

NUMERICAL EXAMPLE

To illustrate the procedure, a simple three period example will be presented and solved, for the case when there is no existing machine at time zero.

Machines may be bought at times 0, 1, 2. The cost of a machine bought at time n is assumed to be

$$C_n = 1{,}000 + 500\, n^2.$$

The discount rate, the fraction instantaneous depreciation at purchase, and the maximum preventive maintenance per period are assumed to be

$$r = 0.06, \qquad \delta = 0.25, \qquad \text{and } U = \$100,$$

respectively.

Let $R_n(t)$ be the net return (net of necessary maintenance) of a machine purchased in period n and operated in period t.

We assume

$$R_0(0) = \$600, \qquad R_1(1) = \$1{,}000, \qquad R_2(2) = \$1{,}100.$$

The incremental return function is given by $R_n(t+1) - R_n(t) \equiv \psi_n(m(t),t)$. We assume $\psi_n(m(t),t) = -d_n + a_n m(t)$, where

$$d_1 = 200, \qquad d_2 = 50, \qquad d_3 = 100, \qquad \text{and } a_n = 0.5 + 0.1\, n^3.$$

This means that the return on a machine purchased in period n goes down by an amount d_n every period if there is no preventive maintenance. This decrease can be offset by an amount proportional to the amount of preventive maintenance.

(Note making this function stationary over time is just done to simplify example.)

Let $S_n(t)$ be the salvage value at t of a machine purchased at n. We assume

$$S_n(n) = (1 - \delta)C_n = 0.75\,[1{,}000 + 500\, n^2].$$

The incremental salvage value function is given by

$$S_n(t+1) - S_n(t) = -l_n C_n + b_n m(t),$$

where

$$l_n = 0.1, \qquad n = 0,1$$

$$= 0.2, \qquad n = 2,$$

and

$$b_n = (0.5 - 0.05 + m_n(t)$$

i.e., the decrease in salvage value is a constant percentage of the purchase price if there is no preventive maintenance. With preventive maintenance, salvage value can be enhanced by a proportional amount.

Let V^*_{nk} be the optimal value of the objective function associated with a machine purchase at n and sold at k. We will now solve for V^*_{nk}, $n = 0, 1, 2$, and $3 \geq k > n$, k an integer.

Before we proceed, we will denote by $W(t/n, k)$, the coefficient of m in Hamiltonian H in (18),

$$W(t/n,k) = -(1+r)^{-t} + a_n \sum_{i=t+1}^{k} (1+r)^{-i} + b_n(1+r)^{-k}.$$

The optimal control is given by (20).

It is noted in passing that

$$Sgn\,[W((t+1)/n,k) - W(t/n,k)] = Sgn\,[r - a_n],$$

implying that

$$\begin{aligned} m^*(t+1) - m^*(t) &\geq 0 \qquad \text{if } (r-a_n) > 0 \\ &= 0 \qquad \text{if } (r-a_n) = 0 \\ &\leq 0 \qquad \text{if } (r-a_n) < 0. \end{aligned}$$

In this example $r - a_n < 0$, implying that if there is a switching in optimal control, it must be from \$100 to \$0.

$n = 0, k = 1$

$$W(0) = -1 + 0.5\,(1.06)^{-1} + 0.5\,(1.06)^{-1} < 0.$$

This implies that

$$m^*(0) = 0.$$

Now,

$$\begin{aligned} &R_0(0) = 600. \\ &R_0(1) = 600 - 200 = 400. \\ &S_0(0) = 0.75 \times 1{,}000 = 750. \\ &S_0(1) = 750 - 0.1 \times 1{,}000 = 650. \\ &V_{01}^* \quad = 600 + 400 \times (1.06)^{-1} - 1{,}000 + 650 \times (1.06)^{-1} = \$590. \end{aligned}$$

Similar calculations can be carried out for other subproblems. We will list these results.

$n = 0, k = 2$

$$\begin{aligned} &m^*(0) = 100, \qquad m^*(1) = 0 \\ &V_{02}^* = 681. \end{aligned}$$

$n = 0, k = 3$

$$m^*(0) = 100, \quad m^*(1) = 100, \quad m^*(2) = 0$$
$$V^*_{03} = 644.$$

$N = 1, k = 2$

$$m^*(1) = 0$$
$$v^*_{12} = 1242.$$

$n = 1, k = 3$

$$m^*(1) = 100, \quad m^*(2) = 0$$
$$V^*_{13} = 1869.$$

$n = 2, k = 3$

$$m^*(2) = 100$$
$$V^*_{23} = 555.$$

Wagner-Whitin Solution of the Entire Problem

With reference to the dynamic programming equation in (30) we have

$$\begin{aligned} f_3 &= 0, \\ f_2 &= V^*_{23} = \$555, \\ f_1 &= \max [V^*_{13}, V^*_{12} + f_2] \\ &= \max [1869, 1242 + 555] \\ &= \$1869, \\ f_0 &= \max [V^*_{03}, V^*_{01} + f_1, V^*_{02} + f_2] \\ &= \max [644, 590 + 1869, 681 + 555] \\ &= \$2{,}459 \end{aligned}$$

Now we can summarize the optimal solution.

First Machine

Purchase at $n = 0$, sale at $k = 1$ optimal preventive maintenance $m^*(0) = 0$.

Second Machine

Purchase at $n = 1$, sale at $k = 3$ optimal preventive maintenance $m^*(1) = 100, m^*(2) = 0$.

The Value of Objective Function $V^* = \$2{,}459$.

REFERENCES

[1] Cannon, M. D., C. D. Cullum, Jr., and E. Polak, *Theory of Optimal Control and Mathematical Programming* (McGraw-Hill, New York, 1970).

[2] Kleindorfer, G. B., P. R. Kleindorfer, and G. L. Thompson, "The Discrete Maximum Principle with Applications to Management Science," *Management Science Research Report No. 137*, GSIA (Pittsburgh, Pa., Carnegie-Mellon University, 1968).

[3] Masse, P., *Optimal Investment Decisions* (Prentice-Hall, Englewood Cliffs, New Jersey, 1962).

[4] Naslund, Bertil, "Simultaneous Determination of Optimal Repair Policy and Service Life," Swedish Journal of Economics *LXVIII*, 63–73 (1966).

[5] Sethi, S. P., "Simultaneous Optimization of Preventive Maintenance and Replacement Policy for Machines," *Management Science Research Report Np.223*, GSIA (Pittsburgh, Pa., Carnegie-Mellon University, 1971).

[6] Sethi, S. P. and G. L. Thompson, "A Conditional Maximum Principle and the Christmas Toy Manufacturers Problem," *GSIA Working Paper*, Carnegie-Mellon University, 1971.

[7] Terborgh, G., *Dynamic Equipment Policy* (McGraw-Hill, New York, 1949).

[8] Thompson, G. L., "Optimal Maintenance Policy and Sale Date of a Machine," Management Science (May 1968), pp. 543–550.

[9] Wagner, H. M., *Principles of Operations Research* (Prentice-Hall, Inc., Englewood Cliffs, N. J., 1969).

[10] Wagner, H. M., and T. M. Whitin, "Dynamic Version of the Economic Lot Size Model," Management Science *5*, pp. 89–96, 1958.

A CLASS OF INVENTORY MODELS WITH CUSTOMER IMPATIENCE

M. J. M. Posner

Department of Industrial Engineering
University of Toronto
Toronto, Canada

and

B. Yansouni

Organization Research & Systems
Bank of Montreal
Montreal, Canada

ABSTRACT

Classical inventory models generally assume either no backlogging of demands or unlimited backlogging. This paper treats the case wherein backlogged customers are willing to wait for a random period of time for service. A broad class of such models is discussed, with a more complete analysis performed on a simple subclass. Steady state equations are derived and solved assuming exponentially distributed interarrival times of customers, order delivery lead times, and customer patience.

INTRODUCTION

The essential difficulties arising in the analysis of classical inventory problems, result from the introduction of the dual uncertainties characterizing customer demand and order procurement lead times. Under assumptions of randomized demand and delivery time lag, it may be impossible or too costly to guarantee that all demands will be consistently satisfied. Failure to satisfy demands results in a financial penalty generally known as the shortage cost. Depending upon the specific model under consideration, shortages and shortage costs may be treated in a multiplicity of fashions. Some models assume that excess demands are satisfied by priority shipments [5], the shortage costs then being reckoned as the additional cost over and above that associated with routine delivery. In others, excess demands are refused altogether [1], implicity implying that management considers efforts to satisfy the customer as entailing costs greater than the value of a lost sale. In some circumstances we may also view demands "turned away" as resulting from the customers' unwillingness to wait for satisfaction.

Where management feels bound to satisfy customers from routine delivery, a waiting line of excess demands may form. In these models it is implicity assumed either that failure to commit the firm to satisfy their customers would result in "goodwill" costs, or that customers are willing to wait indefinitely for service. From a consideration of customer behavior alone, most inventory models treat excess demands as either willing or unwilling to wait for satisfaction.

Hanssman [4] considers these two alternatives by defining the "patient customer" willing to wait indefinitely for service and the "impatient customer" requiring instantaneous satisfaction. Analogously,

Hadley and Whitin [3] referred to the same cases as the backordering and the lost sales case, respectively. The impatient customer case generally arises in a direct customer to seller environment, where the customer leaves for a competing seller if the product is not immediately available. Similarly the patient customer case occurs when there is a sole supplier of the product, as with military procurement, or possibly in mail order shopping.

In reality, however, customer patience generally is not patterned after one of these extremes, but rather assumes an intermediate posture. For example, a person placing an order for a specific truck or automobile model currently out of stock, may wait, but not indefinitely, before deciding to shop elsewhere. Similarly, in any competitive environment where there is some customer loyalty, the firm must ensure that the customer's desire to associate with their goods is not compromised by trying his patience. It might be preferable, for this purchase, that the potential delay in his satisfaction be revealed to him, so that he may decide to purchase elsewhere. Hopefully, cognizant of the realities of stockout, he would thereafter return to his former loyalty. However, if we improperly encourage him to wait for satisfaction, then an untimely delay may encourage him to shop elsewhere. This loss would be realized not only for this purchase, but for future ones as well, if his anger is sufficiently kindled.

Thus, customers are assumed to have limited patience, and will wait for a random period of time for satisfaction, at no additional cost to the company. However, if the customer becomes impatient and leaves before receiving service, the total loss associated with this demand will generally exceed that corresponding to one sale. This specific type of loss may be referred to as the double loss penalty (DLP), since two distinct penalties are involved, namely the actual loss of a sale and the potential loss of future sales expressed as "goodwill." It is therefore the intention of this paper to investigate the implications of introducing customer impatience into inventory models.

THE GENERAL MODEL

The model that will be introduced to characterize customer impatience represents an extension of the classical (T, R) continuous time review and ordering policy and its variations, as described by Arrow, Karlin, and Scarf [1], and Galliher, Morse and Simond [2]. The former demonstrated the applicability of the (T, R) ordering policy applied to the lost sales case. From the point of view of customer behaviour this model assumes that customers are totally impatient. The model involves the specification of two parameters, T and R such that whenever the stock level reaches R an order for T units is placed. Analysis of this model attempts to determine that policy (T^*, R^*) which mimimizes a suitably chosen objective function, incorporating costs of holding, reordering, and shortage or penalty costs due to unsatisfied customers. In arriving at this optimal policy, the countering effects of varying T and R are easily explained: For smaller R shortage risk grows and holding costs decrease, whilst for larger T the frequency of shortages and reorders is reduced, but holding costs rise.

Galliher et al. considered the extension of the equivalent model from the no-backlogging to the infinite-backlogging case. Again from the point of view of customer behaviour, they assume that customers are infinitely patient. If we let $X_t (-\infty < X_t \leq T+R)$ denote the inventory level at time t, then the set of possible ordering levels would be defined as those for which $X_t \in \{R - nT | n = 0,1,2, \ldots\}$. Thus, as more orders remain outstanding and the backlog continues to grow, additional orders are placed at queue length intervals of size T. In this case the backlog may decrease only by virtue of order deliveries.

In these and other similar models, finite backlogging of excess demands is not considered since treatment of customer behaviour has been restricted to the cases of unlimited, or no, patience. With

unlimited patience, formulation of a finite backlogging scheme cannot be easily rationalized, for this mode of customer behaviour presupposes that goodwill is not overly affected. However, when customer impatience is considered, the size of the backlog allowed to form becomes an important variable which management must consider in the determination of a suitable policy.

We will now consider the no-backlogging (T, R) model of Arrow, Karlin, and Scarf as a zero-stage backlogging model, and will then define an extention to a K-stage backlogging model with the following characteristics: An order is placed when $X_t = R$, and again at a later time t' when $X_{t'} = R - T$ while the previous order has not yet been delivered. Similarly, if there are n orders outstanding at some time t, and then $X_{t'} = R - nT$ $(n = 0,1,2, \ldots, K-1)$, an additional order for T units is placed. However, if at time t, K orders are outstanding and $X_t = R - KT$, any additional customer demands arising are turned away. Note that it is implicitly assumed above that $T > R$ so that $R - nT < 0$ for $n \geq 1$, and this restriction is retained throughout.

The general justification for utilizing some K-stage model would be that management feels that customers are willing to wait for a certain period of time for satisfaction, after which their goodwill is placed in jeopardy. It would naturally be assumed that this goodwill is in excess of merely losing a single sale, since otherwise management would enact an infinite-backlogging policy. Should a customer then depart because of impatience, the corresponding loss would be somewhat greater than that reflected in a single lost sale, so that management wishes to utilize the limited customer patience in allowing some queue to form. Nevertheless, it is unwilling to go beyond customer tolerance in desiring satisfaction. If the degree of impatience is weak, that is, if customers are willing to wait for a relatively long time to receive service, then a high shortage risk should be run preferable to a high holding cost, and conversely.

Customer impatience adds a complicating feature to that discussed by Galliher et al. since in their model knowledge of X_t implies with certainty knowledge of the number of orders outstanding at time t. Thus for $n = 1, 2, \ldots, K$, if $R - nT < X_t \leq R - (n-1)T$ it is known that exactly n orders are currently outstanding. However, when customer impatience is considered, and customers may renege from the queue, this one-to-one correspondence is not imparted since the stock level may oscillate about an ordering level. This situation is explained, for example, when the following sequence of events occurs: A demand arises at time t yielding $X_t = R - nT$ $(n = 1,2,3, \ldots, K-1)$, so that $n+1$ orders are now outstanding; a customer becomes impatient and departs from the system; and then a new customer arrives at time t' so that $X_{t'} = R - nT$. It is clearly undesirable to place an additional order at time t . Therefore any mathematical model must keep track of not only the stock level, but also of the number of orders currently outstanding.

Furthermore, an additional complication is encountered with the treatment of multiple outstanding orders. For a realistic class of models, we must ensure that outstanding orders do not cross in time; that is, the orders should arrive in the sequence in which they were placed. Thus, a reasonable model must also keep track of the sequence of elapsed times since the orders were placed.

Clearly, the formulation and solution of the general K-stage model appears to represent a formidable task. It is therefore the intention of this paper to initiate the analysis of finite backlogging models associated with customer impatience by means of a consideration of the single-stage backlogging of customers. The general features of this model will then be investigated to determine the feasibility of investigations of higher order problems.

THE ONE-STAGE MODEL

In the one-stage model an order for T units is placed whenever the inventory level is reduced to R, and the backlog is limited to $T - R$ $(T > R)$. Excess demands are refused, and a maximum of one

order may be outstanding at any time. It is assumed that customer demands arrive singly in a homogeneous Poisson stream of rate λ, and that the order procurement lead time is an exponentially distributed random variable with a mean delivery time of $1/\mu$. Although it may be reasonable for customers to consider the size of a backlog when determining whether to join the queue or to balk and not enter the system, it is assumed for convenience in this paper that the balking phenomenon does not exist. Finally, customer impatience is modelled as an exponentially distributed random variable with mean patience time of $1/\mu$, independent of the queue size. Thus the parameters λ, μ, α, represent, respectively, the mean rates of customer demands, order delivery, and customer departures due to impatience.

The introduction of exponential distributions for each of the three random variables defined enables us to describe a suitable model in the simplest terms. In particular, the use of the exponential distribution as a representation of customer impatience appears intuitively realistic. In the current competitive environment most customers would likely exhibit rather limited loyalty to a product, so that their waiting time for satisfaction could be modelled as a monotonic, quickly decreasing function of time. Thus, most customers would wait only a short while, with few holding fast to their loyalty for an extensive period. Further justification and additional comments regarding the utilization of exponential distributions are presented at the conclusion of the paper.

The procedure to be followed will involve the derivation of the appropriate Chapman-Kolmogorov equations with a restrictive investigation of the steady-state solution only. Although it is possible to treat the transient situation, we are generally concerned with an examination of the system attributes over an infinite time horizon and a comparison with previous models.

If $X_t = k$ we say the system state at time t is k. In the one-stage model, the state space can be viewed as consisting of three distinctive sets, each one of them generated by events of a different nature:

(a) $R < k \leq T + R$: After an order of T units has been delivered the system state must necessarily belong to this region. In addition, no orders are outstanding except in the case $k = R$, wherein an order is placed upon each entry into that state.

(b) $0 \leq k \leq R$: For states in this set an order is currently outstanding; however, these states cannot be reached through the delivery of an order of T units. In both sets, (a) and (b), the impatience factor is not taken into account since no backlog exists.

(c) $-(T-R) \leq k < 0$: For states in this set an order is outstanding, and customer departures due to impatience must be considered.

For convenience set $N = T + R$ and $n = T - R$. For $-n \leq k \leq N$, let P_k^t be the probability that at time t the system state is k. From an elementary consideration of the associated Chapman-Kolmogorv equations, utilizing the unique properties of the exponential distribution, the following differential-difference equations are easily obtained:

$$P_N^{t+dt} = P_N^t(1 - \lambda dt) + P_R^t \mu dt + 0(dt^2),$$

$$P_k^{t+dt} = P_k^t(1 - \lambda dt) + P_{k+1}^t \lambda dt + P_{k-T}^t \mu dt + 0(dt^2), \quad k = R+1, R+2, \ldots, N-1,$$

$$P_R^{t+dt} = P_R^t(1 - \lambda dt - \mu dt) + P_{R+1}^t \lambda dt + P_{R-T}^t \mu dt + 0(dt^2),$$

$$P_k^{t+dt} = P_k^t(1 - \lambda dt - \mu dt) + P_{k+1}^t \lambda dt + 0(dt^2), \quad k = 1, 2, \ldots, R-1,$$

$$P^{t+dt}_{-k} = P^{t}_{-k}(1 - \lambda dt - \mu dt - k\alpha dt) + P^{t}_{-k-1}(k+1)\alpha dt + P^{t}_{-k+1}\lambda dt + 0(dt^2),$$

$$k = 0, 1, \ldots, n-1,$$

$$P^{t+dt}_{-n} = P^{t}_{-n}(1 - \mu dt - n\alpha dt) + P^{t}_{-n+1}\lambda dt + 0(dt^2).$$

We may easily convert the above into a finite set of differential equations. Taking limits as $t \to \infty$, and noting that $\lim_{t\to\infty} P^t_k = P_k$, we obtain the following:

(1) $$-\beta P_N + \gamma P_R = 0;$$

(2) $$-\beta P_k + \beta P_{k+1} + \gamma P_{k-T} = 0, \qquad k = R+1, \ldots, N-1;$$

(3) $$-(\beta+\gamma)P_R + \beta P_{R+1} + \gamma P_{-n} = 0;$$

(4) $$-(\beta+\gamma)P_k + \beta P_{k+1} = 0, \qquad k = 1, 2, \ldots, R-1;$$

(5) $$-(\beta+\gamma+k)P_{-k} + (k+1)P_{-k-1} + \beta P_{-k+1} = 0, \quad k = 0, 1, \ldots, n-1;$$

(6) $$-(\gamma+n)P_{-n} + \beta P_{-n+1} = 0;$$

and

(7) $$\sum_{k=-n}^{N} P_k = 1;$$

where

$$\beta = \lambda/\alpha$$

and

$$\gamma = \mu/\alpha.$$

Equations (1)–(7) clearly possess a unique solution. For moderate sized problems this may be obtained by Gaussian elimination, or other numerical methods. For larger problems, and to determine the effects of parameter variation, an analytic solution is desirable. Thus, in solving the above systems, (5) and (6) are first analyzed to yield P_{-k} $(k = 0, 1, \ldots, n)$ in terms of P_1. Then equations (1)–(4) are considered, obtaining P_k $(k = 1, 2, \ldots, N-1)$ in terms of P_N. At that stage, since all probabilities are expressed in terms of P_N, use of (7) results in the desired solution.

Observe that Equation (6) gives P_{-n+1} in terms of P_n. Then by considering the Equations (5) in reverse order, all P_{-k+1} $(k = 0, 1, \ldots, n-1)$ are obtained by recursion in terms of P_{-n}. These may then be easily adjusted so that all P_{-k} $(k = 0, 1, \ldots, n)$ are expressed in terms of P_1. In addition P_1 as well as P_k $(k = 1, 2, \ldots, R)$ may be solved in terms of P_N by considering Equations (1) and (4). Solving (4) recursively in terms of P_R yields

(8) $$P_k = (\beta/\beta+\gamma)^{R-k}P_R, \qquad k = 1, 2, \ldots, R,$$

$$= (\beta/\beta + \gamma)^{R-k}(\beta/\gamma)P_N,$$

on using (1).

In order to obtain P_k $(k = R+1, \ldots, T)$, consider Equation (2) rewritten as

$$P_k = P_{k+1} + (\gamma/\beta)P_{k-T}, \qquad k = R+1, \ldots, T,$$

which, upon solving recursively in terms of P_{T+1} yields

$$P_k = P_{T+1} + \sum_{j=0}^{k-T} (\gamma/\beta)P_{-j}, \qquad k = R+1, \ldots, T. \tag{9}$$

Also from (2) we find that for $k = T+1, \ldots, N-1$,

$$P_k = P_N + (\gamma/\beta) \sum_{j=k-T}^{R-1} P_j \tag{10}$$

$$= P_N \left[1 + \sum_{j=1}^{N-k} (\beta/\beta+\gamma)^j\right],$$

upon using (8). Thus all P_k $(k = -n, \ldots, N)$ are now expressed in terms of P_N, which is finally determined using (7).

EXAMPLE

Having obtained the stationary solution for the one-stage model, we are now in a position to compare the effects of incorporating customer impatience and DLP associated with the classical zero-stage model of Arrow, Karlin, Scarf. In the model under consideration, a linear cost or return function can be constructed for the stationary process with the following form:

$$\lambda Q = bf_1 - hf_2 - df_3 - sf_4 - rf_5, \tag{11}$$

where

Q = Stationary expected return per customer demand.

b = Net benefit per sale, with f_1 the expected sales rate.

h = Cost of holding an item in inventory per unit of time, with f_2 the expected stock size.

d = Double loss penalty (DLP) suffered per impatient departure, with f_3 the expected rate of impatient departures.

s = Stockout penalty suffered per demand refused service, with f_4 the expected rate of refusal of customer demands.

r = Reordering cost, with f_5 the expected rate of entering the reorder level.

From elementary considerations, it is easily determined that

$$f_5 = \lambda P_{R+1} + \mu P_{-n}$$

(12) $$f_4 = \lambda P_{-n}$$

(13) $$f_2 = \sum_1^n kP_k.$$

Since the expected number of backlogged customers becoming impatient per unit time is $k\alpha$ when the queue size is k, it follows that

(14) $$f_3 = \alpha \sum_1^n kP_{-k}.$$

Similarly, we find that

(15) $$f_1 = \lambda \sum_1^N P_k + \mu \sum_1^n kP_{-k},$$

where the first term is the expected sales rate from stock on hand, and the second is the expected number of backlogging customers satisfied per unit time by virtue of order deliveries.

For convenience, introduce the parameters

$$\theta = \mu/\lambda$$
$$\xi = \alpha/\lambda$$

(so that in previous derivations, $\gamma = \theta/\xi$, $\beta = 1/\xi$). Therefore, using Equations (11)–(15) an expression for Q may be written as

(16) $$Q = b\sum_1^N P_k + (b\theta - d\alpha)\sum_1^n kP_{-k} - (h/\lambda)\sum_1^N kP_k - (s + \theta r)P_{-n} - rP_{R+1}.$$

We shall consider a hypothetical example wherein a recently formed company selling trucks is eager to promote its business operations and to develop a good reputation. We will assume that the management has obtained and will rely on the following cost figures. The selling price per unit is \$4,280 and the cost of purchase from the manufacturer is \$3,880 per unit. The net benefit per sale is 5.8 percent of the selling price or $b = \$250$. This figure takes into account the usual expenditures from monthly payment of debts, advertising, etc., as might arise for recently formed companies. The holding cost is $h = \$9.5$ per unit per 5-day period, including the effects of costs of tied-up capital, rental of parking space, and insurance. It is further assumed that the effective cost of turning away a demand is $s = \$40$, reflecting a minor dissatisfaction or irritation should a customer be refused service. Note that this cost is not what is referred to as the DLP, since demands are turned away irrespective of the customers' willingness to wait. Finally we will assume that customer demands form a homogeneous Poisson process of rate $\lambda = 1.4$ units per 5-day period. This data is substituted into Equation (16) which is then optimized with respect to the parameters T and R, expressed for varying values of θ, ξ, and DLP. Graphical representation of these results may be found in Figure 1, with the optimal returns for the zero-stage (T, R) policies superimposed.

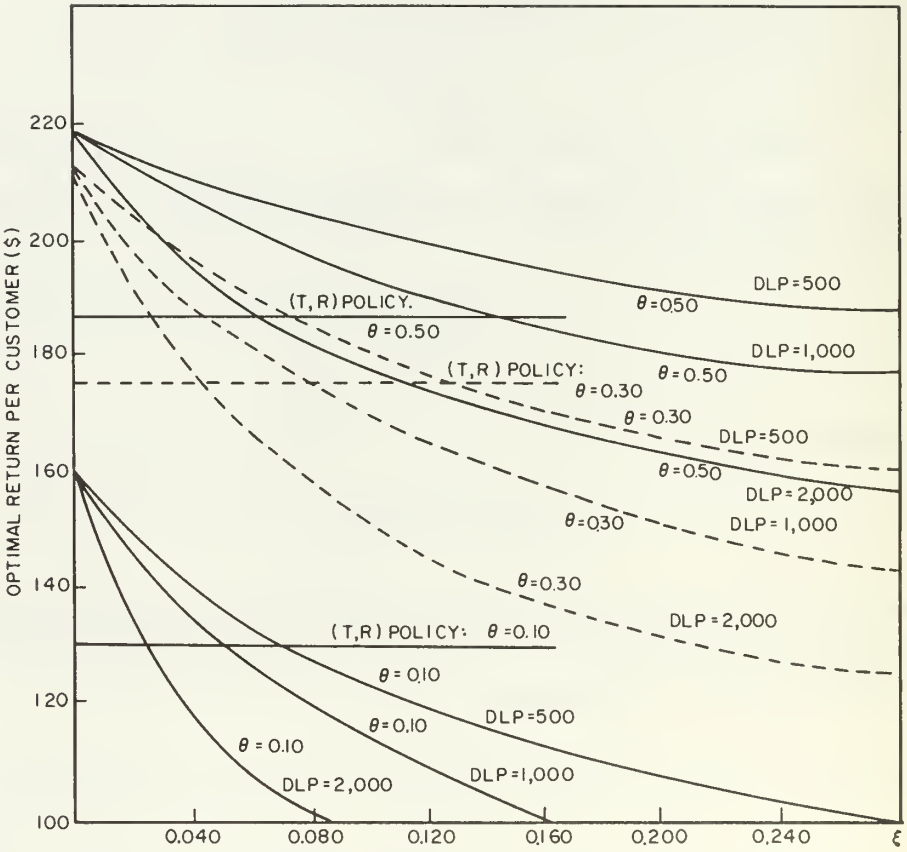


FIGURE 1. Optimal return per customer in terms of ξ for various values of θ and DLP.

It is quite clear that as the impatience factor α increases, reflected in increasing ξ, the optimal return per customer decreases. Indeed as $\xi \to \infty$, all customers applying during shortages leave instantaneously, and no queue is formed. Since there is no penalty attached to departures when $DLP = 0$, the model behaves exactly as the (T, R) policy in which the stockout losses due to impatient customers would be ignored. The horizontal lines represent the optimal returns generated by zero-stage ordering policies for various values of θ. Observe that for each value assumed by the θ and DLP, we may define a critical value for ξ, say ξ_c, such that for all $\xi < \xi_c$, the optimal return obtained with the one-stage model is greater than that for the zero-stage model while the contrary holds true if $\xi > \xi_c$. This information is graphically demonstrated in Figure 2. Note that for a given value for DLP, ξ increases with increasing θ. This is expected since as θ increases, the shortage risks decrease and fewer customers are willing to wait for service. Therefore the impatience factor becomes less critical, and we can afford to accept the commitment to satisfy more customers without incurring additional loss.

CONCLUSIONS

Implicit in the above treatment is the question of quantizing the DLP. Determination of a good measure of DLP involves the separate aspect of designing a suitable experiment. However, as in the described example, only approximate measures of DLP are needed to yield results which present qualitative information as to the need and conditions wherein policies other than one of the extremities should be considered.

Although the results outlined referred to a specific one-stage class of models, it is clear from the

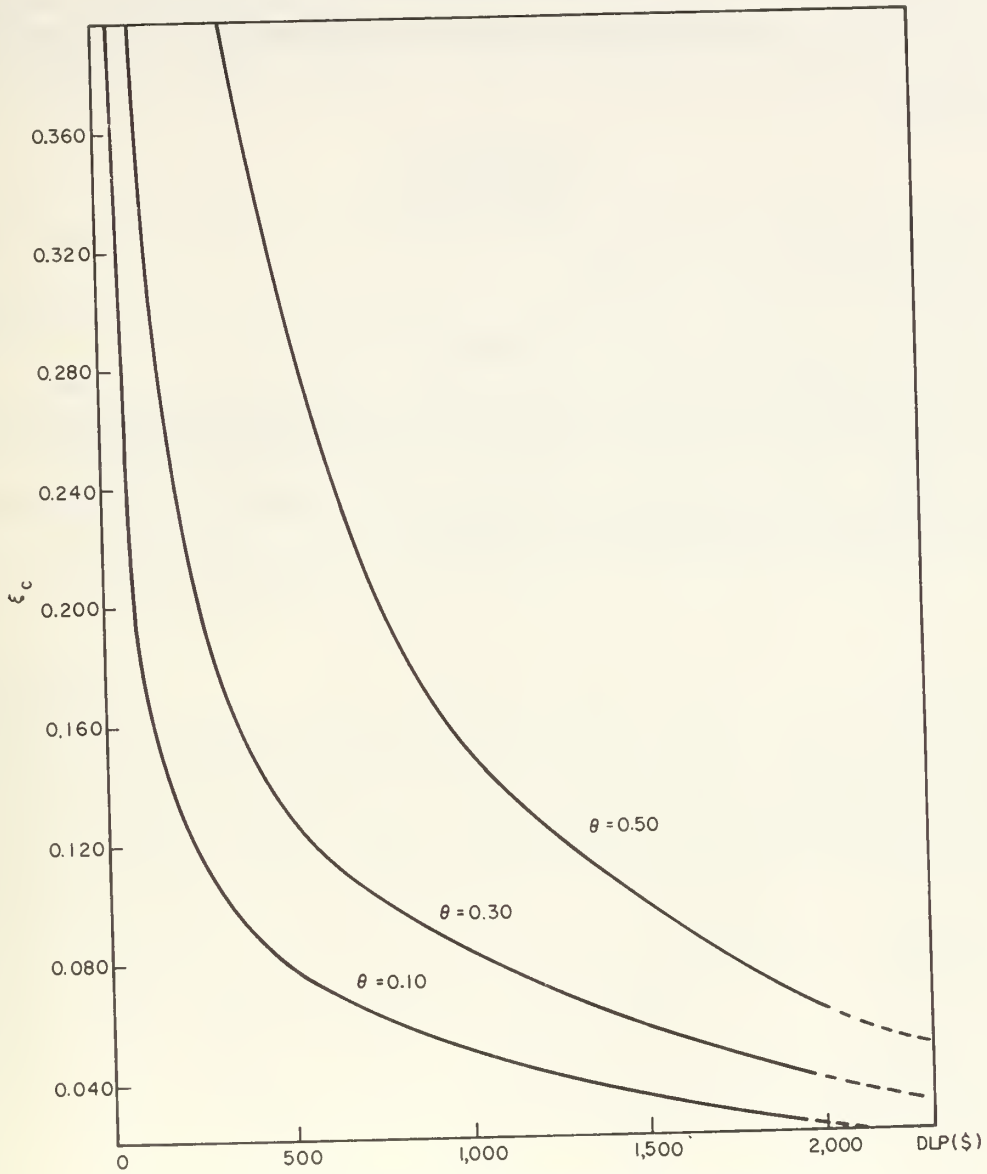


FIGURE 2. Critical values of ξ expressed in terms of DLP.

interesting nature of the results that further efforts should be directed towards analysis of multiple stage models, in which the number of stages also becomes a parameter subject to optimization. In addition, other models might be investigated including a consideration of balking phenomena, wherein customer arrivals may be deterred from entering the system if the backlog is too large. We may also consider customer impatience in situations where the degree of this impatience depends upon the backlog size. Although in general the resulting sets of equations would be difficult to solve, in many instances the number of such equations is fairly small, and may be solved simply by Gaussian elimination.

Extensions of these models to the treatment of nonexponential forms for the impatience characteristics and the delivery time lag might also prove fruitful. We should expect that the use of the random exponential distributions throughout (coefficient of variation = 1) has resulted in rather conservative estimates of the return. The lesser randomness of the class of Erlang distributions (coefficient of varia-

tion < 1) should lead to improved values for the return. However, for purposes of comparing alternative policies, the use of exponential distributions appears to provide a simple yet meaningful mode for decisionmaking.

REFERENCES

[1] Arrow, K. J., S. Karlin, and H. Scarf, *Studies in the Mathematical Theories of Inventory and Production* (Stanford Univ. Press, 1951).

[2] Galliher, H. P., P. M. Morse, and M. Simond, "Dynamics of Two Classes of Continuous-Review Inventory Systems", Operations Research **7**, 362–384 (1959).

[3] Hadley, G., and T. Whitin, *Analysis for Inventory Systems* (Prentice-Hall, N.J., 1963).

[4] Hanssman, F., Operations Research in Production and Inventory Control (John Wiley and Sons, N.Y., 1962).

[5] Wright, G. W., "Optimal Ordering Policies for Inventory Systems with Emergency Ordering," Operational Research Quarterly **20**, 111–123 (1969).

SYSTEMS APPROACH TO THE MULTISTAGE MANUFACTURING CONNECTED-UNIT SITUATION

Sedat Sami Ercan

Roosevelt University
Chicago, Illinois

ABSTRACT

This paper investigates the relationships among incoming quality limit (*AIQL*), manufacturing quality level in each stage of production, and outgoing quality limit (*AOQL*), for a multistage manufacturing connected-unit situation from the systems cost-effectiveness point of view. An objective function is developed and the formulated problem is solved using discrete optimization technique.

The study is restricted to the development of single sampling plans where inspection is by attribute. Lot size and sample size are fixed for incoming and outgoing inspection for each stage of manufacturing.

INTRODUCTION

The quality control engineer, like other staff and managers, must be concerned with the systems aspects of the enterprise. Because of his responsibility for quality he must, therefore, be directly concerned with economic and systems aspects of quality control function of management. The one important fact that makes any systems oriented economic analysis of a situation hazardous is that the conditions of supply, demand and costs are dynamic and therefore, the parameters of today's analysis are not necessarily valid for tomorrow's decisions.

Several researchers have contributed to an understanding of systems-affect on the economics of quality control. Their contributions can be summarized as follows:

On the economic selection of a quality level, Shahnazarian [10] presented a model in which multicomponents were used in a single assembly and introduced the concept of opportunity loss when an outgoing quality level is arbitrarily selected without considering the quality levels of incoming components.

Smith [11, 12] and Martin [8, 9] studied the economics of sampling inspection. They introduced the probability associated with cost and extended the result to derive an optimal sampling plan without taking into consideration the effect of the incoming quality level and assuming the occurence of costs as random variables.

Enell [3] discussed the selection of *AQL* or *AOQL* by using a breakeven equation based on unit cost of acceptance, cost of inspecting a piece, cost of repairing or replacing a defective, and the cost of rejecting the lot under the assumption of constant probability.

Towards the design of optimal sampling plans, Johns and Guthrie [7] and Hald [6] presented an asymptotic solution for finding the sample size and acceptance number which minimizes the expected risk assuming that each lot is formed as a sequence of Bernoulli events from a process with a constant probability, "p," of producing defectives.

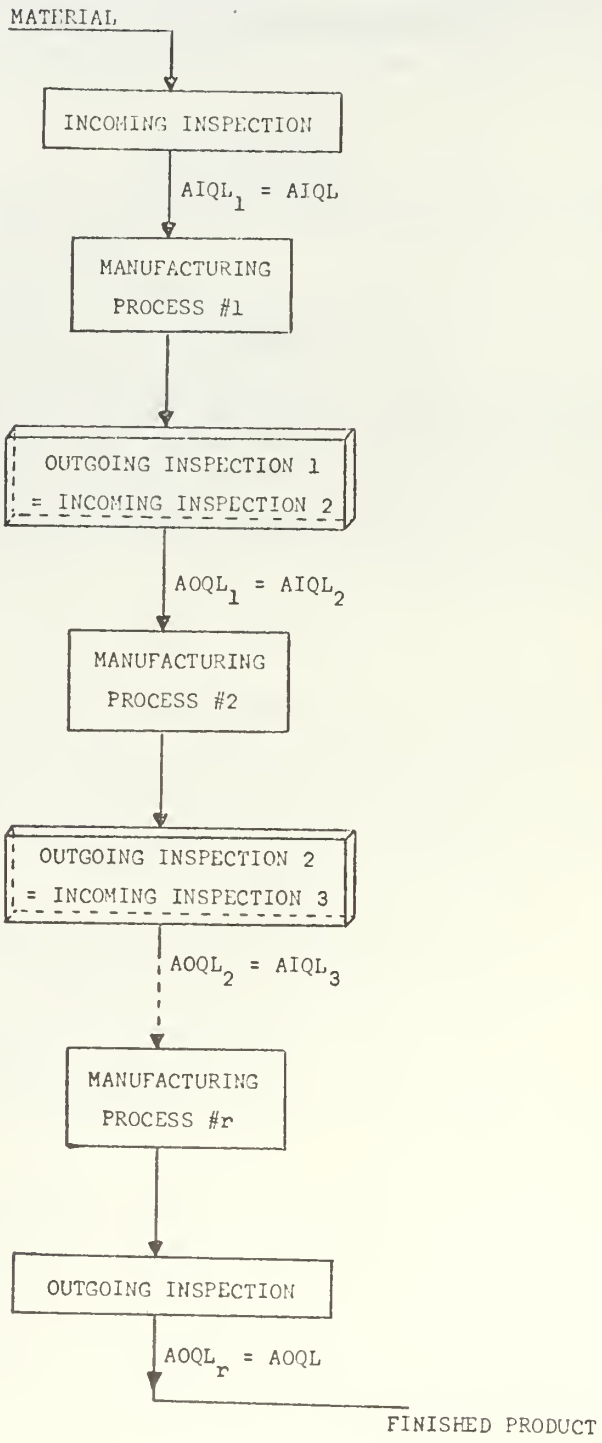


FIGURE 1. Systems representation of multistage manufacturing *AIQL–AOQL* relation of a connected-unit situation.

Interrelated sampling plans are studied by Beightler and Mitten [1] using the transition matrix of probability between the stages of operation in a sequential manner. Their conclusion is that "the sequences of sampling plans which are truly optimal for an entire system cannot be found by treating each station in isolation from all others, since savings may be made for some stations at the expense of increased costs at later stages."

The use of systems oriented economics in the design of quality control systems is not widespread because the concepts proposed so far in this field of study are limited in their treatments. The concept proposed in this article recognizes the interrelations among incoming quality limit, manufacturing quality level and outgoing quality limit. A model for systems cost-affected single sampling plans for a multistage manufacturing connected-unit situation is developed. A single sampling plan is obtained using discrete optimization for the total expected loss function.

The manufacturing situation to be considered in this paper may be described as follows: raw material after incoming inspection will go through several manufacturing processes and the connected-unit will be inspected after each manufacturing process. This will be called outgoing inspection, and will be the incoming inspection of the following manufacturing process. The system of Multistage Manufacturing Connected-Unit situation then is composed of incoming inspection, manufacturing processes and outgoing inspection as shown in Figure 1.

AVERAGE INCOMING QUALITY LIMIT *(AIQL)*—AVERAGE OUTGOING QUALITY LIMIT *(AOQL)* QUALITY CONTROL PROBLEMS

AIQL—AOQL Quality Control problems may be classified as:

1. Either *AIQL* or *AOQL* is not fixed and W is constant. The problem then is to determine the one which is not fixed, and the problem in this case will be called a deterministic *AIQL—AOQL* quality control problem.

2. Either *AIQL* or *AOQL* is not fixed and W is not constant; that is, the number of defectives created through manufacturing has a probability function. The problem then is to find the optimum combination of *AIQL* and *AOQL* for the possible values of W with the minimum cost of the system. This problem will be called an *AIQL—AOQL* quality control problem under uncertainty which will be treated in this paper.

In general for any single-stage manufacturing, *AOQL* can be written as the algebraic sum of *AIQL* and W of the same stage [4],

$$AOQL = AIQL + W, \tag{1}$$

where *AOQL* is the Average Outgoing Quality Limit and *AIQL* is the Average Incoming Quality Limit, and W, a stochastic variable, is the function of the number of defectives created from the manufacturing process, that is, the amount of increase or decrease in *AOQL* due to manufacturing process characteristics. Thus,

$$W = g(x), \tag{2}$$

where x is the number of defectives created from manufacturing process.

Let $p_N^n(x)$ be a discrete manufacturing function, producing x number of defectives when the

process function is $f(p)$, where $f(p)$, is a probability density function for the process fraction defective p; when p is constant using binomial, then the probability of getting x number of defectives in a sample size n is given as follows [7]:

$$p_N^n(x) = \begin{cases} \binom{n}{x} \int_0^1 p^x (1-p)^{n-x} f(p)\,dp & \text{for } x \geq 0 \\ 0 & \text{elsewhere.} \end{cases} \tag{3}$$

Thus, $p_N^n(x)$ is a function of n and N. Furthermore, it is also a probability function [5]. This property of $p_N^n(x)$ is used in deriving the condition for systems cost-affected single sampling plans.

FORMULATION OF THE MULTISTAGE MANUFACTURING CONNECTED-UNIT SITUATION OBJECTIVE FUNCTION

Total expected loss function for the Multistage Manufacturing Connected-Unit Situation can be written as follows [4]:

$$L = \sum_{i=1}^{r} \left[\sum_{x=0}^{c^i} k_1^i (c^i - x) p_N^{n^i}(x) + \sum_{x=c^i+1}^{n^i} k_2^i (x - c^i) p_N^{n^i}(x) + k_3^i n^i \right] + k_4 n^r, \tag{4}$$

where

$\sum_{x=0}^{c^i} k_1^i (c^i - x) p_N^{n^i}(x)$ is the expected opportunity loss for manufacturing stage i,

$\sum_{x=c^i+1}^{n^i} k_2^i (x - c^i) p_N^{n^i}(x)$ is the expected repair loss for manufacturing stage i,

$k_3^i n^i$ is the loss due to incoming inspection cost for manufacturing stage i,

$k_4 n^r$ is the loss due to outgoing inspection cost for the last stage of manufacturing,

and where the unit is connected where connectedness is as defined by Debreu [2] "a subset S of R^m if it is not the union of two non-empty, disjoint subsets closed in S." In other words, a set S is connected if it cannot be partitioned into two non-empty subsets closed in S. Practically speaking, the unit is made of one piece,

k_1^i is the opportunity loss, at manufacturing stage i, in dollars per unit, due to underestimation and/or overestimation of $AOQL$, such that

$$AOQL_i = AIQL_{i+1} \qquad \text{for } i = 1, 2, 3, \ldots, r-1, \text{ and} \tag{5}$$

$$AOQL_i = AOQL \text{ for } i = \text{r}.$$

k_2^i is the average cost of replacement or repair of a defective, at manufacturing stage i, in dollars per unit,

k_3^i is the incoming inspection cost, in dollars per unit, at manufacturing stage i without regard to quality where $i = 1, 2, 3, \ldots, r$.

k_4 is the outgoing inspection cost, in dollars per unit, at manufacturing stage r without regard to quality,

n^i is the sample size at manufacturing stage i, and it is assumed to be unique for the system,

c^i is the difference between outgoing and incoming acceptance numbers at manufacturing stage i, $c^i = C_0^i - C_I^i$, where $i = 1, 2, 3, \ldots, r$. Furthermore, $c^i = C_I^{i+1} - C_I^i$ where $C_0^i = C_I^{i+1}$.

The Multistage Manufacturing Connected-Unit Situation *AIQL–AOQL* problem from the systems cost-effectiveness point of view can be formulated as follows: *AOQL* is a function of C_0 and *AIQL* is a function of C_I and because of well-ordering principles of C_0 and C_I, in order to devise a single sampling plan, it is enough to determine c^i for specified *AOQL* which in turn determines the *AIQL* of the system. Then the problem to obtain systems cost-affected single sampling plans for this situation can be written explicitly in optimization form as

$$\text{(6)} \qquad \min_{c^i} L = \sum_{i=1}^{r} \min_{c^i} \left[\sum_{x=0}^{c^i} k_1^i (c^i - x) p_N^n(x) + \sum_{x=c^i+1}^{n} k_2^i (x - c^i) p_N^n(x) \right],$$

where c^i is an integro-valued function.

ANALYTICAL SOLUTION TO THE MULTISTAGE MANUFACTURING CONNECTED-UNIT SITUATION

Equation (6) can be minimized using discrete optimization technique. The necessary conditions to minimize Equation (6) are

$$\text{(7)} \qquad \Delta L^{c^i(t)} \geq 0 \qquad \forall i,$$

a forward difference of L on $c^i(t)$, and

$$\text{(8)} \qquad \nabla L^{c^i(t)} \leq 0 \qquad \forall i,$$

a backward difference of L on $c^i(t)$; where t is a dummy variable. Equation (7) is evaluated, and the result is given as

$$\text{(9)} \qquad [\Delta c^i(t)\ (k_1^i + k_2^i) \sum_{x=0}^{c^i} p_N^n(x) - k_2^i] \geq 0 \qquad \forall i,$$

where

$$\Delta c^i(t) = c^i(t+1) - c^i(t) = 1.$$

Thus, Equation (9) reduces to

$$\text{(10)} \qquad \sum_{x=0}^{c^i} p_N^n(x) \geq \frac{k_2^i}{k_1^i + k_2^i} \qquad \forall i.$$

Similarly, from Equation (8), Equation (11) can be obtained as follows:

$$\frac{k_2^i}{k_1^i + k_2^i} \geq \sum_{x=0}^{c^i - 1} p_N^n(x) \qquad \forall i. \tag{11}$$

Combining Equations (10) and (11), Equation (12) is written as

$$\sum_{x=0}^{c^i - 1} p_N^n(x) \leq \frac{k_2^i}{k_1^i + k_2^i} \leq \sum_{x=0}^{c^i} p_N^n(x) \qquad \forall i, \tag{12}$$

which is a well-known formula in inventory systems.

The sufficient conditions for minimizing Equation (6) can be written as $\Delta^2 L^{c^i(t)} > 0 \; \forall t$ which could be proven from Equation (12) that it has one or at most two solutions. It is also known that Equation (6) is a convex function [5], and $\Delta^2 L^{c^i(t)} > 0 \; \forall t$ is satisfied and need not be considered further. Therefore, Equation (12) will give the global minimum of Equation (6), which in turn will achieve a set of $\{n, C_j^i, C_0\}$ single sampling plans for k_j^i, where $i = 1, 2, \ldots, r$, and $j = 1, 2, 3, 4$.

Example: A plant produces a "one-component product" in three production stages with the manufacturing function of

$$p_N^n(x) = \begin{cases} \dfrac{e^{-2} 2^x}{x!} & x \geq 0 \\ 0 & \text{elsewhere,} \end{cases}$$

for each stage of manufacturing.

Find the systems cost-affected single sampling plans when $N = 500$, $n = 10$, and $AOQL$ is specified as 0.4465, and when the plant uses the following cost matrix:

Manufacturing stage $= i$	k_1^i	k_2^i	k_3^i	k_4
1	\$1.75	\$1.25	\$0.50	—
2	1.00	2.00	1.00	—
3	0.50	1.00	0.75	\$1.50.

Solution: The given $p_N^n(x)$ and cumulative $p_N^n(x)$, for different values of x are tabulated as follows:

x	$p_N^n(x)$	$\sum_x p_N^n(x)$
0	0.1353	0.1353
1	0.2707	0.4060
2	0.2707	0.6767
3	0.1804	0.8571
4	0.0902	0.9473
5	0.0361	0.9834
6	0.0120	0.9954
7	0.0034	0.9988
8	0.0009	0.9997
9	0.0002	0.9999
10	0.0001	1.0000.

Utilizing the given cost matrix and using Equations (4), (5), and (12) for $i = 1, 2, 3$, and for $n = 10$, $N = 500$, the systems cost-affected single sampling plan is obtained from Table 1 for the specified $AOQL$ as follows:

$$C_0 = 7 \qquad n = 10 \qquad AOQL = 0.4465$$

$$C_1 = 2 \qquad n = 10 \qquad AIQL = 0.134358$$

and the minimal total cost for the system is \$42.2009.

TABLE 1. *Systems Cost-Affected Single Sampling Plans For Three-Stage Manufacturing*

First Stage

C_0^1	$C_1^1 = C_1$	$AOQL_1 = AIQL_2$	$AIQL_1$	Minimum cost
1	0	0.0823984	0.0360542	\$6.6489
2	1	0.134358	0.0823984	6.6489
[a] 3	[b] 2	0.190316	[b] 0.134358	6.6489
4	3	0.249312	0.190316	6.6489
5	4	0.310464	0.249312	6.6489
6	5	0.370201	0.310464	6.6489
7	6	0.4465	0.370201	6.6489

Second Stage

C_0^2	C_1^2	$AOQL_2 = AIQL_3$	$AIQL_2$	Minimum cost
2	0	0.134358	0.0360542	\$12.0347
3	1	0.190316	0.0823984	12.0347
4	2	0.249312	0.134358	12.0347
[a] 5	[a] 3	0.310464	0.190316	12.0347
6	4	0.370201	0.249312	12.0347
7	5	0.4465	0.310464	12.0347

Third Stage

C_0	C_1^3	$AOQL$	$AIQL_3$	Minimum cost
2	0	0.134358	0.0360542	\$23.5173
3	1	0.190316	0.0823984	23.5173
4	2	0.249312	0.134358	23.5173
5	3	0.310464	0.190316	23.5173
6	4	0.370201	0.249312	23.5173
[b] 7	[a] 5	[b] 0.4465	0.310464	23.5173

[a] Indicates the dummy (in between stages) acceptance numbers for the optimal single sampling plan.

[b] Indicates $AIQL$ and specified $AOQL$ and acceptance numbers for the optimal single sampling plan.

CONCLUSION

This paper recognizes the interrelations among incoming quality limit (*AIQL*), manufacturing quality level, and outgoing quality limit (*AOQL*). An analytical solution has been proposed using discrete optimization for the systems oriented total expected loss function.

The study is restricted to the development of single sampling plans where inspection is by attribute. Lot size and sample size are fixed for each stage of manufacturing. A single sampling plan has been derived using a numerical example utilizing the proposed analytical solution.

REFERENCES

[1] Beightler, C. S. and L. G. Mitten, "Design of an Optimal Sequence of Interrelated Sampling Plans," J. Am. Stat. Assoc. *59*, 96–104 (Mar. 1964).

[2] Debreu, G., *Theory of Value—An Axiomatic Analysis of Economic Equilibrium*, Cowles Foundation for Research in Economics at Yale University (John Wiley and Sons, Inc., New York, 1959).

[3] Enell, J. W., "Which Sampling Plan Should I Choose?" Industrial Quality Control *10*, 96 (May 1954).

[4] Ercan, S. S., "Average Incoming—Outgoing Quality Levels Problems," Unpublished Master Thesis, Illinois Institute of Technology, Department of Industrial Engineering (1969).

[5] Ercan, S. S., "Cost-Affected Single Sampling Plans Subject to AIQL—AOQL Equality Constraints," Ph.D. Dissertation, Illinois Institute of Technology, Department of Industrial Engineering (1970).

[6] Hald, A., "Determination of Single Sampling Attribute Plans with Given Producer's Consumer's Risk," Technometrics *9*, 401–415 (Aug. 1967).

[7] Johns, M. G., Jr. and D. Guthrie, Jr., "Bayes Acceptance Sampling Procedures for Large Lots," Ann. Math. Stat. *XXX*, 896–925 (1959).

[8] Martin, C. A., "Determination of an Optimum Acceptance Sampling Plan Based on Cost," Proceedings of Middle Atlantic ASQC Conference (1962).

[9] Martin, C. A., "The Cost Breakeven Point in Attribute Sampling," Industrial Quality Control *21*, 135–143 (Sept. 1964).

[10] Shahnazarian, R. E., "Selection of Economical Quality Levels for Multicomponent Assemblies," Industrial Quality Control *23*, 173–183 (Oct. 1965).

[11] Smith, B. E., "Some Economic Aspects of Quality Control," *Technical Report 53*, Applied Mathematics and Statistics Laboratories, Stanford University (July 3, 1961).

[12] Smith, B. E., "The Economics of Sampling Inspection," Industrial Quality Control *22*, 453–458 (Mar. 1965).

ON THE ECONOMIC APPLICATION OF AIRLIFT TO PRODUCT DISTRIBUTION AND ITS IMPACT ON INVENTORY LEVELS

Donald Gross

The George Washington University
Washington, D.C.

and

A. Soriano

Air Reduction Company
Murray Hill, New Jersey

ABSTRACT

In many resupply situations, the decisionmaker has the option of "purchasing" faster replenishment leadtimes. For example, a premium may be paid for delivery by parcel post rather than slower but less expensive delivery by railway express. It may be economically advantageous to pay shipment premiums for faster leadtimes when considering the possible cost reductions in pipeline (on-order) inventory and safety stock levels. This paper develops a decision rule which, for any given item, will indicate whether it is economically advantageous to purchase a faster leadtime. The general methodology is then applied to a peacetime military resupply operation involving several million items, each requiring a decision as to whether the item should be shipped by air or sea.

I. INTRODUCTION

The problem treated in this paper deals with the development of a decision rule for determining when it is economically advantageous to purchase faster leadtimes in product resupply systems. The inventory policy is assumed to be a periodic review, (s, S) policy; that is, inventory is reviewed on a periodic basis (once per week, one per month, etc.) and if the on-hand plus on-order (pipeline) inventory is at or below a level s, an order is placed to bring the level to S. Complete backordering of demand is assumed so that it is possible for on-hand inventory to be negative, indicating a backorder condition. The backorders are then filled when the next shipment arrives.

Reducing the leadtime can lead to cost savings in two ways. First, pipeline inventory is reduced. Second, a smaller safety stock is required to insure a given service criterion such as a limit on the probability of a stockout or the percent of demands backordered. In some situations, it may be only the latter cost savings which are of importance, since the supplier may be outside the system and if a bill is not rendered until goods are received, in effect the supplier bears the pipeline costs. We assume, here, however, that the pipeline costs are borne by the resupply system. It is an easy matter to set these to zero when this is not the case.

Both demand and leadtime are assumed to be random variables. Purchasing premium leadtimes amounts to essentially being able to reduce the mean or the mean and variance of the leadtime distribution. A reduction in mean leadtime directly reduces the pipeline inventory. A reduction in mean or mean and variance of the leadtime distribution allows for smaller safety stock levels without adversely

affecting the preset criterion of system service. The particular service criterion used in this analysis is the percent of demand which is backordered, that is, the percent of customers who cannot be immediately satisfied from on-shelf inventory and must wait for their requests to be filled. The replenishment system is depicted in Figure 1. The leadtime is represented by T (a random variable) and r denotes the period between reviews. It is assumed that r is fixed by management policy.

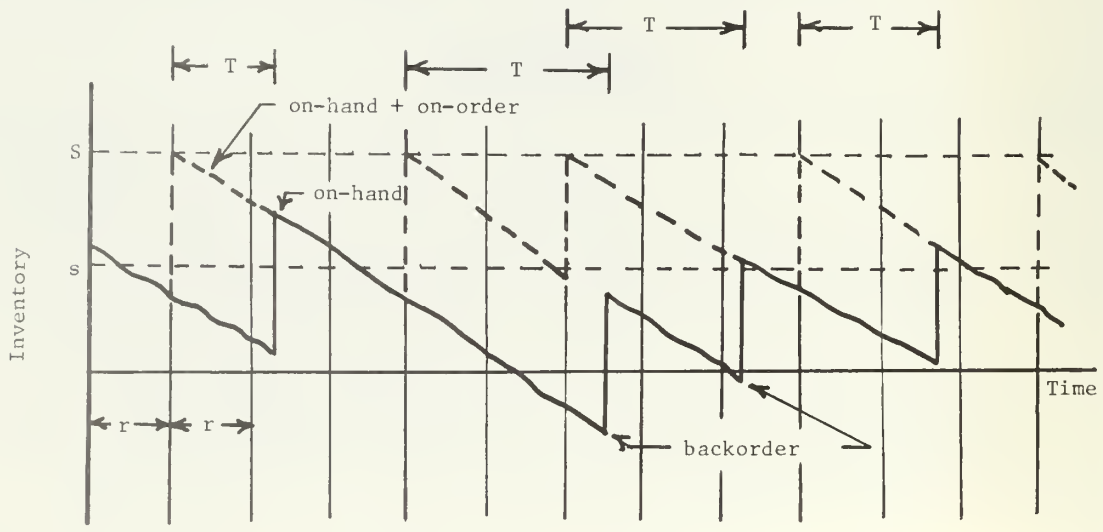


FIGURE 1.

No attempt is made here to determine the optimal values of s and S. We assume these are fixed for some current resupply mode. The effect to be studied is the economic impact of being able to reduce mean leadtime (or the mean and variance) by paying a premium for a faster transportation mode.

II. DECISION RULE

We introduce the following notation for developing the decision rule.

X = demand per unit time
$f(x)$ = probability distribution of X
λ = mean demand = $E[X]$
T_j = leadtime, mode j
$g_j(t)$ = probability distribution of T_j
τ_j = $E[T_j]$
$I_{p,j}$ = average pipeline inventory using mode j
$I_{h,j}$ = average on-shelf inventory using mode j
A_j = percent demand filled directly from on-shelf inventory using mode j
c_p = carrying cost for pipeline inventory
c_h = carrying cost for on-shelf inventory (\$/unit/unit-time)
$c_{T,j}$ = cost of resupply transportation per unit using mode j

Since mode i is preferred to mode j if the cost reductions achievable by carrying less pipeline and on-shelf inventory are greater than the extra transportation costs incurred, the decision rule can be written as follows.

Choose mode i over mode j for each item where

$$c_p(I_{p,i} - I_{p,j}) + c_h(I_{h,i} - I_{h,j}) + c_{\tau,i} - c_{\tau,j} < 0, \tag{1}$$

subject to

$$A_i = A_j. \tag{2}$$

Equation (1) can be rewritten as

$$c_p \Delta I_{p;i,j} + c_h \Delta I_{h;i,j} + \Delta c_{\tau;i,j} < 0, \tag{3}$$

where

$$\Delta I_{p;i,j} = I_{p,i} - I_{p,j},$$

$$\Delta I_{h;i,j} = I_{h,i} - I_{h,j},$$

and

$$\Delta c_{\tau;i,j} = c_{\tau,i} - c_{\tau,j}$$

We first turn our attention to $I_{p,j}$, the expected on-order (pipeline) inventory. We use an expected value analysis to show that $I_{p,j} = \lambda\tau_j$ as follows. When an order is placed for Q units, on the average it is on order for a time τ. The average number of orders placed per unit time is λ/Q, where Q is the average order quantity. This must be so, since all demand is eventually satisfied. Thus, on the average, λ/Q times per unit time, we have Q units on order for time τ. If, for example, our basic time unit is days then the average on-order inventory is given by ($Q\tau$ unit-days per order) (λ/Q orders per day), or $\lambda\tau$ units. Therefore,

$$I_{p,j} = \lambda\tau_j \tag{4}$$

and

$$\Delta I_{p;i,j} = \lambda(\tau_j - \tau_i) = \lambda\Delta\tau_{i,j}\ . \tag{5}$$

Unfortunately, the difference in average on-shelf inventory, $\Delta I_{h;i,j}$, is not so easy to derive. This involves the constraint as given by Equation (2). An exact analytical development of I_h when both demand and leadtimes are stochastic is quite difficult and in some cases impossible. Some models can be developed if one is willing to assume certain distributions for demand and leadtimes. Since leadtimes are stochastic, a conflict arises between allowing leadtimes to be independent random variables and not having orders cross, that is, a subsequent order received prior to one placed earlier. Both order crossing and non independence of leadtimes present great analytical difficulty (see [2], pp. 200–204; 285–287).

Rather than make any assumptions concerning demand for the particular application which follows, a Monte Carlo simulation model of a periodic (s, S) inventory system was developed and used for the purpose of assessing $\Delta I_{h,i,j}$ for a given A (see [1]).

The calculation of $\Delta c_{\tau;i,j}$ depends entirely on the particular modes of transportation used in the resupply process so that no general expression can be obtained. We, therefore, at this time turn to a particular application of the decision rule given by Equation (3) subject to the constraint of Equation (2).

III. A MILITARY PEACETIME RESUPPLY SYSTEM

The application presented here involves a peacetime military resupply system. Overseas bases must be periodically resupplied from depots in the United States. Generally, in peacetime, routine resupply is done by ships (sealift). However, airlifting supplies provides for faster leadtimes, usually at an added shipment cost. Conceptually then, the decision rule developed above can be applied to each item to see if there is an advantage to shipping it by air rather than by sea. There are over two million items, however, in this resupply system. Practical considerations made it necessary to deal with classes of items, rather than single items. Similar items (similar with respect to average usage, weight, volume, cost, etc.) were grouped into a "commodity class." A commodity class was then treated as a single item, having the characteristics of the average of all items in the class. Thus, although the analysis appears to be on an item by item basis, it, in reality, is on a group of items so that a decision favoring airlift over sealift for a particular commodity class applies to all items in that class. The error introduced by grouping items depends on how diverse are the characteristics of the items in the class. One must trade off the number of analyses required involving the decision rule against the errors introduced by large groupings.

For each class of items going from a stateside depot to an overseas location, the decision rule is applied to a composite item of the class exhibiting the average characteristic of the class. The resupply inventory policy is a periodic (s, S) policy as follows. The inventory status of each item was reviewed on a weekly basis. The trigger point, s, was set at the average pipeline inventory plus "one month of supply."* Denoting mean weekly demand by λ, we have then

$$s = 4.3\,\lambda + I_p. \tag{6}$$

The upper level value, S, was set at one month's supply above s, that is

$$\begin{aligned} S &= s + 4.3\,\lambda \\ &= 8.6\,\lambda + I_p. \end{aligned} \tag{7}$$ †

For sealift, the mean leadtime was estimated at 13 weeks while for airlift, the mean leadtime was estimated at 2 weeks. The reduction in pipeline inventory can be easily calculated from Equation (5) to be

$$\Delta I_{p,a,s} = -11\,\lambda, \tag{8}$$

where "a" represents the airlift mode and "s" represents the sealift mode.

---

*Month of supply is a common term in use in the military and refers to the expected monthly usage, in this case 4.3λ.

†No claim is made that either s, or S is optimal. These were set by current regulations in effect at the time of the study.

To estimate $\Delta I_{h;a,s}$ the simulation model was employed. Although good estimates of mean leadtime and mean demand were available, estimates of other factors, such as the distribution shapes and even the variances were not available. A variety of cases were run on the simulator, using s and S as given by Equations (6) and (7), respectively. Poisson and normal distributions were used for demand, each with a variety of coefficients of variation. For leadtime simulation, normal, uniform, exponential, and constant (deterministic) distributions were used, again with a variety of coefficients of variation where possible (for example, the exponential distribution always has a coefficient of variation of 1.0). For explicit details on the simulation and cases considered, see [1].

From past history and field experience, it was known that the service criterion as defined above and denoted by "A" was approximately 95 percent.* Since current policy prior to the study was to sealift all items, the simulation was run with a mean leadtime of 13. Cases which gave a markedly different A from 95 percent were discarded as not realistic. For example, using the exponential distribution for leadtime gave an A of around 85 percent and was considered a poor choice for describing the probability mechanism governing leadtimes. Likewise, deterministic leadtimes gave an A near 100 percent, a situation which was also considered unrealistic. Had estimates of leadtime variation been available, these cases undoubtedly could have been ruled out a priori.

The summary of the simulation results are presented in Tables I and II. Table I presents results where s was first set according to established policy rules, that is, according to Equation (6) for the sealift case ($\tau_s = 13$). The simulation provided an estimate for average on-shelf inventory under this condition for each of the cases run. Then τ was reduced to two ($\tau_a = 2$) and the average on-shelf inventory obtained in order to achieve the same A value as that for sealift for each of the simulations cases. From this, $\Delta I_{h;a,s}$ was computed for each case and a summary presented in the table.

TABLE I. *Realistic Cases Run Under Current Policy Setting for* s
($\tau_s = 13$; $s = 17.3\ \lambda$ from Equation (6))

A(%)			$\Delta I_{h;a,s}/\lambda$		
Average	Maximum	Minimum	Average	Maximum	Minimum
97	99	96	−3.3	−3.4	−3.2

TABLE II *Realistic Cases Run, Setting* s *to yield an* A = 95 percent

A(%)	$\Delta I_{h;a,s}/\lambda$		
	Average	Maximum	Minimum
95	−2.7	−3.3	−2.0

In Table II, current procedure was not used for setting s, but the s used was that necessary to achieve an A of 95 percent. Again, the results are summarized for the realistic cases run. A conservative estimate of the potential reduction in average on-shelf inventory in utilizing airlift over sealift would be two and one-half weeks of supply. Thus, for ΔI_h we have

$$\Delta I_{h;a,s} = -2.5\ \lambda. \tag{9}$$

*A constant service level for all items imputes a high stockout cost to high cost items and a low stockout cost to low cost items. Further, this implies equal essentiality of all items. This may not be optimal, but for this case was prescribed policy.

It is necessary, in order to calculate the dollar savings in inventory to determine a value for c_p and c_h. Although it is possible that the carrying charge between pipeline and on-shelf inventory could differ (loss, obsolescence, etc. may differ) if the major portion of the carrying charge is due to the cash investment in inventory, these factors should be approximately equal. For this study, it was assumed that $c_p = c_h$. Further, the carrying cost generally can be considered proportional to the dollar value of the item so that

$$c_p = c_h = k_1 c, \tag{10}$$

where k_1 is the constant of proportionality and c is the unit cost. Various studies (see Reference [3] for example) have shown that k_1 varies between 0.15 and 0.30. Three values of k_1 were used in this application: 0.20, 0.25, and 0.30. For each value, the decision rule was applied to all commodity classes.

It remains now only to estimate the difference in transportation costs, $\Delta c_{\tau;a,s}$. Consider the following notation:

m_j = average resupply distance in miles via mode j
w_j = weight of the items in tons when transportation mode j is used
c_j = cost of transporting one ton of the item one mile, i.e., \$/ton-mile, via mode j.

It might, at first seem strange that the weight of an item depends on transportation mode. This occurs because of different packaging requirements; packaging for sealift is usually heavier than packaging for airlift.* Also mileage depends on transportation mode, air mileage generally being shorter than sea mileage. It is assumed that the storage areas to which goods are shipped are sufficiently close to sea and airports so as to make ground transportation considerations insignificant. The difference in transportation costs can then be written as

$$\Delta c_{\tau;a,s} = (c_a w_a m_a - c_s w_s m_s)\lambda. \tag{11}$$

Generally, for all items, the ratio of w_a to w_s will be approximately the same and, from a given depot to storage area, the ratio of m_a to m_s is independent of the particular item. Thus letting

$$k_2 = m_s/m_a$$

$$k_3 = w_s/w_a,$$

we can write Equation (11) as

$$\Delta c_{\tau;a,s} = w_a m_a \lambda (c_a - k_2 k_3 c_s). \tag{12}$$

The factors k_2 and k_3 can be readily estimated by sampling some typical situations. By using Equations (8), (9), (10), and (12) in Equation (3), the decision rule becomes:

"If

$$-13.5 k_1 c + w_a m_a (c_a - k_2 k_3 c_s) < 0, \tag{13}$$

ship by air rather than sea."

*Volume considerations are not explicitly considered here; however volume could be handled in a similar manner.

Note that the first term is negative and the second will generally be positive since in almost all situations, $c_a > k_2 k_3 c_s$. The first term represents the cost reduction due to reduced inventory possible when airlifting while the second term represents the added transportation cost of airlifting. Note also that λ, the mean demand rate, factors out of the decision rule. Expensive items (large c) which are light in weight (small w_a) will be likely candidates for airlifting since the left hand side of Equation (13) will tend to be negative. This certainly agrees with intuition.

After eliminating obvious items from airlift potential (size or weight prohibitive) the remaining items were grouped into about 800 commodity classes. To estimate the total savings achievable by airlifting the commodity classes so indicated by the decision rule, we merely have to multiply the left hand side of of (13) by λ, where λ is the total mean demand rate of all items in the class, and sum over all classes for which airlift is indicated. Figure 2 shows how potential dollar saving varies with $c_a - c_s$, the cost differential between airlift and sealift, for the three k_1 factors considered.

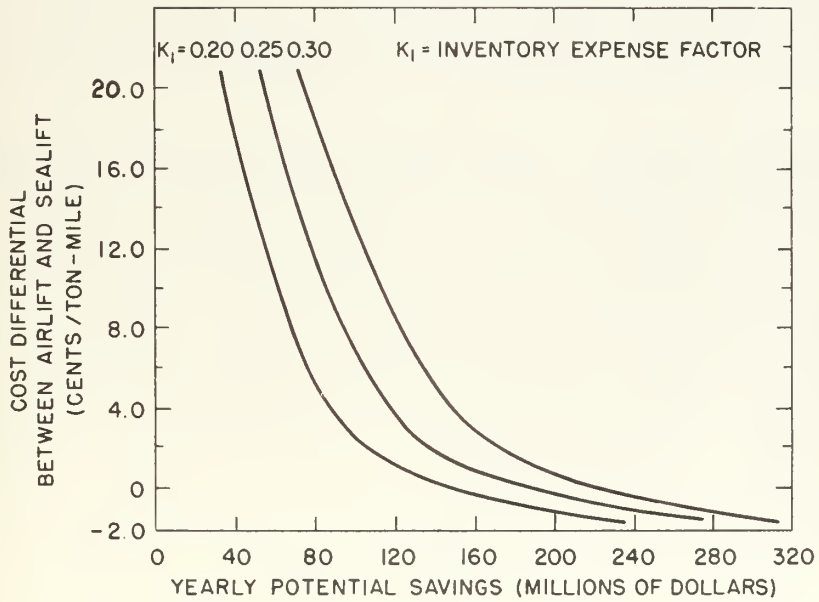


FIGURE 2.

No effort was made here to analyze any costs or savings resulting during the transition period where items are switched from sealift to airlift. This is not to say the transition problem is not important; however, only the steady-state situation was considered here.

REFERENCES

[1] Gross, D. and A. Soriano, "The Effect of Reducing Leadtime on Inventory Levels-Simulation Analysis," Management Science **16**, B–61 — B–76 (1969).

[2] Hadley, G. and T. M. Whitin, *Analysis of Inventory Systems* (Prentice Hall, New Jersey, 1963).

[3] Harbridge House, Inc., "The Costs of Supply Operations," Economic Inventory Policy Report #2 (1959).

BAYESIAN PRIOR DISTRIBUTIONS FOR MULTI-COMPONENT SYSTEMS

J. B. Parker

UK Atomic Energy Authority
Aldermaston, Berks
England

1. INTRODUCTION

The problem of assessing the reliability of a system composed of components each of which has been subjected to go, no-go testing has been approached in many ways during the last two decades. Methods for determining lower confidence limits for a system composed of independent components in series fall naturally into three groups. First there are frequentist methods, examples of which are given in [2], [5] and [13]. Second there are Bayesian methods, the prior distributions of the component reliabilities being assumed to have some explicit mathematical forms [3] and [11]. Third there are asymptotic methods [1] and [6], which are attractive when the number of trials of the components is large.

A full review of the background to these and other matters (notably the time-to-failure problem) is found in [7]. In this and a succeeding article [8] the emphasis is directed on searching for "optimum priors" in the sense that the confidence bounds on system reliability obtained from their use agrees with those implied by classical frequentist methods. Our aim is less ambitious; it is to make a very limited exploration of the idea of choosing the *system* rather than the *component* prior as the basic commodity of prior information. This is an attitude of mind also considered by Mastran [9]. There are two main motivations for this idea. Firstly results achieved by considering the basic commodity of prior information to be the component prior are frequently very pessimistic because of the widespread use of a rectangular component prior; secondly any reliability statement is clearly less sensitive to assumptions made about the system as a whole than to those made about each and every component. As against this, it is admittedly true that assigning a prior to a system rather than to each of its components is open to considerable philosophical objections.

2. THE BAYES APPROACH

The use of Bayes' Theorem for reliability estimation procedures [3] and [10] is so widely known as to require no special explanation. By defining a posterior distribution for a component reliability, p of the form

$$b(p)dp = w(p)L(p)dp \Big/ \int_0^1 w(p)L(p)dp,$$

where w is an (assumed) prior probability function, and $L(p)$ is the likelihood of the trial [$L(p) = {}^nC_m p^{n-m}(1-p)^m$] one can compute p_b such that, for example,

$$\int_{p_b}^{1} b(p)dp = 0.90.$$

More generally, for a system whose configuration is known, we can express the *system* reliability as a known function, $V(p)$, of the several conponent p's. Simulation methods for finding the frequency distribution of $V(p)$ by randomly sampling from the several component posterior distributions have been developed.

Such methods naturally hinge on the choice of w, and this is an area where there has been a good deal of discussion. In the absence of any quantitative information before test, prescribing a functional form for w in (0, 1) is not easy, and calls for judgment based on one's ideas about the performance of the component before test. Many workers have been attracted by the two parameter Beta distribution and without doubt this approach is motivated by the attractive mathematical properties of a Beta distribution when used in conjunction with a binomial likelihood [3]. Such properties are too well known to require description here. It seems to us that it is really no easier to estimate the parameters of the prior Beta distribution, in the absence of previous quantitative test information, than to assert the true mathematical form of one's prior attitudes.

3. CONSIDERATIONS BEHIND THE CHOICE OF A PRIOR

In many practical situations deductions about the reliability of a system are sensitive to the form of the priors assumed for the components. If the $w(p)$ of section 2 is the Beta distribution

$$w(p) = p^a(1-p)^b/B(a+1, b+1),$$

and the trials data consists of n tests, m of which failed, the a posteriori mean value of $1-p$, the component unreliability, is $(m+b+1)/(a+b+n+2)$. For small values of m, this is clearly highly sensitive to the choice of b. Now if the system consists of k components in series, with k moderately large, there will probably be several components having low values of m; if this were not the case the overall system would not have a very high reliability.

If $a = b = 0$ we have a rectangular prior, and the mean value of $1-p$ is $(m+1)/(n+2)$ which result is well known. That this choice of component prior may on occasion be widely misleading can be illustrated by considering a six component series system. Each component is tested 100 times; on five occasions no failures are observed, and on one occasion a single failure. Using a rectangular prior for each component, the mean system reliability is $(101/102)^5(100/102) = 0.93$ approx.

Now suppose 100 systems had been built up out of the 100 sets of components. Ninety-nine systems would have succeeded and one would have failed. By using the same formula the mean system reliability is calculated as 0.98 and moreover the lower 90-percent confidence limit is 0.962. The different result arises since, if we have several events occurring in series, each with uniform prior probability, the prior probability of the combined event (i.e., the system) is not uniform.

We therefore choose as our basic philisophical concept the *system* prior rather than the *component* prior, and in section 4 we shall demonstrate the mathematical linkage between the two. Of course a system prior is even more difficult to allocate than a component prior, though most engineers responsible for designing a system would assert that it was more likely to work than not. Even this rudimentary knowledge suffices to rebut the concept of a uniform component prior, for this results in a hopelessly pessimistic system prior, as we shall show in section 4.

There are other reasons for making the system prior our basic concept. The first one is that system reliability assessments, if sensitive to the choice of component prior (and a case has been illustrated where the sensitivity is marked) will be less so to a choice of system prior; this is simply because (in a series system of k components) we base our prior knowledge on one function rather than k separate ones. A second reason is the following. Given a system, and leaving aside all questions of economy, an inspector could carry out many different strategies. At one extreme he could construct 100 systems and test the lot, calculating a 90-percent reliability figure either using a suitable Bayes prior or else by using a "confidence" approach based on the work of Neyman and Pearson [4]. But he could also view his system as a small number of subsystems in series, each of which he could test on a go, no-go basis. Again he could break the system down even more finely. We believe it to be a self-evident truth that the same system reliability conclusions should be drawn no matter whether the system is conceived and tested as a unit, as a combination of a small number of subsystems, or as a larger combination of a numerous number of components.

This can only arise if the prior for the system is the same in all cases. Correspondingly a prior for a subsystem (a constituent of a k-series system, low k) must be more favorable—i.e., give higher a priori probability to the larger p and a prior for a component (a constituent of a k-series system, high k) a good deal more favorable. This after all is just applied common sense; if a designer is fairly confident that a subsystem will function, he must be much more confident in the performance of its several constituents.

A knowledge of each component prior defines the system prior, which (in the case of a series system) can be obtained by straightforward mathematics. By actually doing this computation, and examining the end product for physical realism, we have a method for assessing whether our original choice of component priors was or was not practical. It is not true of course that a system prior defines the component priors uniquely except in the particular case where all components are deemed to have the same prior.

It is reasonable to expect, however, that different component prior allocations, subject to the system prior being held constant, will not materially affect our judgments on system reliability. This is a matter to which we shall return in due course.

A particular choice for a system prior is a rectangular distribution. This implies that the priors for the components comprising the system will, when expressed as distributions, gravitate towards high values of p, to an extent that is less or greater according to whether k is low or high. If these skewed component priors were used in reliability assessment work (e.g., computer simulation studies) we should certainly get, at any rate for some classes of problem, more realistic estimates of system reliability results (e.g., 90-percent reliability figures) than if a rectangular distribution had been chosen as prior for each component.

4. THE RELATIONSHIP BETWEEN SYSTEM PRIOR AND COMPONENT PRIOR

In the context of a series system composed of k components the relationship between the component priors $w_i(p)\,(i=1, 2, \ldots, k)$ and the system prior $W(p)$ is defined as follows [12];

$$W(p)=\frac{1}{2\pi i}\int_{c-i\infty}^{c+i\infty} M^*(s)p^{-s}ds$$

$$M^*(s) = \prod_{i=1}^{k} M_i(s)$$

$$M_i(s) = \int_0^1 p^{s-1} w_i(p)\,dp, \text{ the Mellin transform of } w_i.$$

Here c is any suitable number such that $W(p)$ exists.

By use of this mechanism the following specimen table, of which the first [12] and last [11] lines have been derived by Springer and Thompson, can be drawn up.

Component prior	$M(s)$	$M^*(s)$	System prior
1. Rectangular	$1/s$	s^{-k}	$[\Gamma(k)]^{-1}(-\log p)^{k-1}$
2. $[\Gamma(1/k)]^{-1}(-\log p)^{(k-1)/k}$	$s^{-1/k}$	$1/s$	Rectangular
3. ~ log normal	$\exp -\left(s\mu + \frac{1}{2}s^2\sigma^2\right)$	$\exp -\left[s(k\mu) + \frac{1}{2}s^2(\sigma\sqrt{k})^2\right]$ ~ log normal	
4. Beta (a, b)	$\lambda(s) = \frac{\Gamma(a+b+2)\Gamma(s+a)}{\Gamma(a+1)\Gamma(s+a+b+1)}$	$[\lambda(s)]^k$	—

We now discuss this table. The system prior corresponding to a rectangular component prior has ($k > 1$) a pole at $p = 0$, vanishes at $p = 1$, and has a mean of 2^{-k}, a phenomenon mentioned by Mastran [9]. By any practical standard, this is a ludicrous system prior.

For a rectangular system prior, the (common) component prior is found to be a function which vanishes at $p = 0$, has a pole at $p = 1$, and has a mean of $2^{-1/k}$. This is certainly not the type of prior distribution advocated by reliability practitioners but we believe that its use in operational studies would not lead to nonsensical conclusions. Next, if we assert that the negative logarithm of p is distributed normally with mean μ and standard deviation σ (this assertion is of course only an approximation since $-\log p$ falls in the range $(0, \infty)$) as the component prior, the system prior has also an approximate log-normal distribution with a higher mean $-\log p$ (and therefore with a lower actual mean) and a correspondingly greater spread. This result is perhaps more easily seen when it is remembered that the logarithm of the system reliability is simply the sum of the logarithms of the several component reliabilities.

A discussion of what is essentially the convolution of several Beta distributions, with different parameters, is given by Springer and Thompson in [11], to which we refer the reader for a statement, in the form of a series, of the system prior. The system prior is complex and is not a member of the Beta class. The rectangular component prior is of course merely a special case of this result ($a = b = 0$).

As remarked earlier, in section 3, a table that is an extension to that given above could be of real service to the system reliability practitioner. At the very least he should examine the repercussions of his original choice of component priors by assessing the merits of the computed system prior. But we would like to go further than this, and, with some temerity, do so, though on a conjectural basis. While the above table has been drawn up on the assumption that each component has the same prior, it is possible to allocate, in some cases quite straightforwardly, different component priors to different components, the system prior remaining unchanged. Next, while acknowledging that the component

prior described in line 2 of the table appears quite fatuous engineering-wise, we yet believe that its operational use, say in simulation studies, will give realistic results, and cite the results described in the next section to buttress our belief. Thirdly a log normal type of prior has an attraction that the Beta prior has not; it permits of ready scrutiny of the resultant system prior. The log-normal prior, like the Beta prior, is a two parameter family and although it enables only unimodal distributions to be approximated (unlike the Beta system) this restriction is not likely to be operationally important.

5. CALCULATIONS

The material assembled in the last two sections will be illustrated with reference to the system described in section 3, a six-component series system, for which the test results are 100 successes out of 100 trials for five components, and 99 successes out of 100 trials for the remaining component. This example is deliberately simple, all the failures occurring in one component, and the numbers on test being the same for all six components. For this example, as we have already noted in section 3, we may regard the data as 100 complete systems (99 of which are successful), and the lower 90-percent reliability confidence limit is 0.962. This is a true confidence limit [4] and independent of Bayesian considerations. By choosing a rectangular prior for the system the mean reliability is 0.98 and the optimum classical lower 90-percent reliability bound of 0.962 is achieved. Asymptotic methods [1], which give 0.961 for the 90-percent reliability figure, are of course satisfactory for this example.

We now evaluate the mean reliability of the system by considering the several constituents component by component, using as priors $\Gamma(1/k_i)^{-1}(-\log p)^{-(k_i-1)/k_i}(i = 1, 2 \ldots 6)$ with the sole restriction that $\sum_{i=1}^{6} 1/k_i = 1$. Using the results of section 4, we observed that this choice of component priors would result in a system prior that was rectangular, the particular choice $k_i = 6(i = 1, 2 \ldots 6)$ corresponding to identical component priors. Using these priors as the w of section 2 and retaining general values n for the sample size (here 100) and k for the number of components (here 6) elementary calculation shows that the mean a posteriori component reliability is $[(n+1)/(n+2)]^{1/k_i}$ in the event that no failures were observed and $[(n+1)^{-1/k_i} - (n+2)^{-1/k_i}]/[n^{-1/k_i} - (n+1)^{-1/k_i}]$ if one failure were observed. The product of the k means, assuming that the trials were all successful for $k-1$ of the components, while the ith component had just one failure, is clearly

$$M = \left[\prod_{\substack{j=1 \\ j\neq i}}^{k} [(n+1)/(n+2)]^{1/k_i}\right]\left[\frac{(n+1)^{-1/k_i} - (n+2)^{-1/k_i}}{n^{-1/k_i} - (n+1)^{-1/k_i}}\right].$$

Remembering that $\sum_{j=1}^{k} 1/k_j = 1$ and expanding in powers of $1/n$ we find

$$M = \frac{n+1}{n+2}\left[\left(\frac{n+2}{n+1}\right)^{1/k_i} - 1\right]\left[\left(\frac{n+1}{n}\right)^{1/k_i} - 1\right]^{-1}$$

$$= n/(n+2) + \left(1 - \frac{1}{k_i}\right)/2n^2.$$

The first term for M is $(n = 100)$ the mean reliability of a system which, when tested 100 times, exhibits just one failure, a rectangular prior distribution being assumed for the system. The second

term is two orders of magnitude lower. To an excellent order of approximation, therefore, we find the same mean reliability by considering the separate components as was adduced by considering the system as a unit, and this is true, however, the system prior (in this case a rectangular prior) is allocated to the contributory components.

Lastly, we actually constructed a Monte Carlo program which sampled from the six appropriate $b(p)$ distributions of section 2, using, for prior, $w(p)=\left[\Gamma\left(\frac{1}{6}\right)\right]^{-1}(-\log p)^{-5/6}$. The details of this program were quite sophisticated, but we do not discuss these here. The objective was to calculate the 90-percent reliability figure. The computed result was 0.964 which is consistent, within sampling error, with the result 0.962 achieved by viewing the data in the form of a hundred complete systems.

6. DISCUSSION

To advocate using a uniform system prior rather than a uniform component prior in reliability assessment problems is certainly not the object of our studies, though we do claim that the former attitude gives a more reasonable approximation to the truth. What we do advocate is that in any practical example component priors should not be assessed, however carefully, in isolation; methods of calculating the corresponding system prior are, at any rate for series systems, available (section 4) and should be used as a check on realism. In this connection, and only in this connection, certain families of priors enjoy suitable combinatorial properties, of which the log normal (i.e., the prior distribution of the negative logarithm of reliability is taken to be Gaussian) is perhaps the most attractive example.

7. ACKNOWLEDGEMENT

While the author assumes sole responsibility for the views adopted in this paper, a substantial part of the work arises from the exchange of ideas with many other people of this Establishment. Also I am particularly grateful for the comments and guidance of Dr. W. R. Buckland. The concept of a log normal prior first came to my attention, though not in the application of this article, through some unpublished work of Colonel L. W. F. Lukis, R.M. Most important, I am extremely grateful to the referee for many thought provoking and wide ranging comments, most of which have been included in the article.

REFERENCES

[1] Bol'shev, L. N., and E. A. Loginov, "Interval Estimates in the Presence of Nuisance Parameters," Theory of Probability and its Applications Vol. XI No. 1, 82–94.

[2] Connor, W. S., and W. T. Wells, "Simulating Tests of a System from Tests of Its Components," Proceedings of the 8th National Symposium of Reliability and Quality Control, 14–16 (1962).

[3] Fox, B. L., "A Bayesian Approach to Reliability Assessment," NASA Memorandum RM/5084–NASA (Aug. 1966).

[4] Kendall, M. G., and A. Stuart, *The Advanced Theory of Statistics* (Griffin, 1961), vol. 2, chap. 20, pp. 103–105.

[5] Lloyd, D. K., and M. Lipow, *Reliability—Management, Methods and Mathematics* (Prentice Hall International, 1962). (The reference is to the Lindstrom and Madden method).

[6] Madansky, A., "Approximate Confidence Limits for the Reliability of Series and Parallel Systems," Technometrics 7, 495–503 (1965).

[7] Mann, Nancy R., "Computer-Aided Selection of Prior Distributions for Generating Monte Carlo Confidence Bounds on System Reliability," Nav. Res. Log. Quart. *17*, 41–54 (1970).

[8] Mann, Nancy R., and F. E. Grubbs, "Approximately Optimum Confidence Bounds on Series System Reliability for Exponential Time to Fail Data," Biometrika, *59*, 191–204 (April 1972).

[9] Mastran, D. V., "A Bayesian Approach for Assessing the Reliability of Air Force Re-Entry Systems," Proceedings of the ASME Reliability and Maintainability Symposium 380–383 (1968).

[10] Schick, G. J., and R. J. Prior, "Reliability and Confidence of Serially Connected Systems," Proceedings of 3rd Space Congress, Cocoa Beach, Florida, 352–360 (March 1966).

[11] Springer, M. D., and W. E. Thompson, "Confidence Limits for the Product of Binomial Parameters," Biometrika p. 611 (Dec. 1966).

[12] Springer, M. D., and W. E. Thompson, "The Distribution of Products of Independent Random Variables," J. Siam Appl. Math. *14*, 511–526 (1966).

[13] Weaver, A. K., "The Determination of the Lower Confidence Limit on the Reliability of a Serial System," NATO Conference on Reliability, Turin (1969).

EFFECT OF SWITCH-OVER DEVICES ON RELIABILITY OF A STANDBY COMPLEX SYSTEM

Pritam Das

Ministry of Defence
New Delhi, India

ABSTRACT

The behaviour of a standby redundant complex system under imperfect switch-over devices with general waiting and repair time distributions for both types of components has been studied. The Laplace transform of various state probabilities have been derived. A particular case has also been discussed.

INTRODUCTION

In the past, some efforts have been made by some authors [2–4] to study the behavior of a standby complex system assuming that sensing and switch-over devices are perfect. Later Parkash [5] relaxed this assumption and evaluated the effect of an imperfect switch-over device on the reliability of the complex system. In his model, he assumed the waiting time distributions for failed components to be exponential. However, in practical situations, it may be observed that waiting time distribution may follow any other distribution.

Keeping this in mind, it is proposed to relax this assumption, so as to make the investigation more practical. The model under consideration has been studied under the following assumptions:

1. The probability of successful operation of a failure sensing and switch-over device is b, which means that its failure probability is $(1-b)$. The repair of this sensing and switching device follows an exponential distribution with mean ν^{-1}.

2. Class L_1 consists of k number of identical components connected in standby redundancy.

3. Class L_2 consists of n components connected in series.

4. The system fails when all the k sub components of L_1 fail, or when any one of the components of L_2 fails, or when the sensing and switch-over device fails.

5. In class L_1, failure follows an exponential distribution with mean a^{-1}, whereas its waiting and repair times follow a general distribution with probability densities $S(X)$ and $\phi(y)$, respectively.

6. In class L_2, failure of the ith component follows exponential distribution with mean λ_i^{-1}, and its waiting and repair time follow general distributions with probability densities $\psi_i(z)$ and $k_i(u)$, $(i=1,2,\ldots,n)$, respectively.

The various states of the model mentioned above can be represented with the help of the flow diagram given in Figure (1).

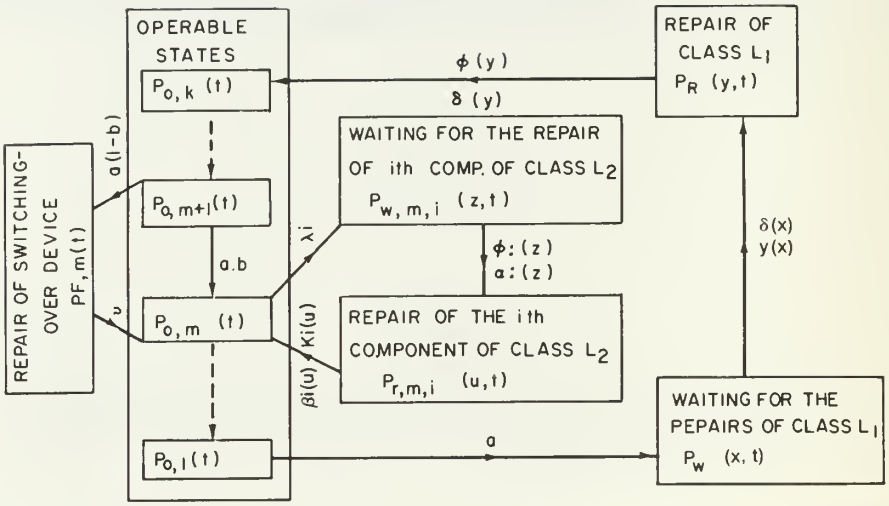


FIGURE 1. Flow diagram.

FORMULATION OF THE MODEL

Define,

$P_{o,m}(t)=$ the probability that at time t, the system is in operable state while m subcomponents of L_1 are in working state, $o \leq m \leq k$.

$P_{F,m}(t)=$ the probability that at time t, the system is in the failed state, due to the imperfect working of the switch-over device, while m number of subcomponents of L_1 are in working order.

$P_{w,m,i}(z,t)=$ the probability that at time t, the system is waiting for the repair of the ith component of L_2 and the elapsed waiting time lies between $(Z, Z+\Delta)$ and m subcomponents of L_1 are in order.

$P_W(x,t)\,\Delta \equiv$ the probability that at time t, the system is waiting for repair of all the K subcomponents of class L_1 and the elapsed waiting time lies in the interval $(x, x+\Delta)$.

$P_{r,m,i}(u,t)\,\Delta \equiv$ the probability that at time t, the system which had failed due to the failure of the ith component of L_2 is being repaired, and the elapsed repair time lies between $(u, u+\Delta)$ and m subcomponents of L_1 are in working order.

$P_R(y,t)\,\Delta \equiv$ the probability that at time t, the system is under repair of class L_1 and the elapsed repair time lies between $(y, y+\Delta)$

From elementary probability considerations, we obtain,

$$\left(\frac{\partial}{\partial t}+\lambda+a\right)P_{o,m}(t)=abP_{o,m}(t)+\nu P_{F,m}(t)+\sum_{i=1}^{n}\int_0^{\infty}P_{r,m,i}(u,t)\beta_i(u)du, \qquad 1\leq m\leq k-1. \tag{1}$$

$$\left(\frac{\partial}{\partial t}+\lambda+a\right)P_{o,k}(t)=\sum_{i=1}^{n}\int_0^{\infty}P_{r,k,i}(u,t)\beta_i(u)du+\int_0^{\infty}P_R(y,t)\delta(y)dy. \tag{2}$$

$$\left(\frac{\partial}{\partial t}+\nu\right)P_{F,m}(t)=a(1-b)P_{o,m+1}(t), \qquad 1\leq m\leq k-1. \tag{3}$$

$$\left(\frac{\partial}{\partial t}+\alpha_i(z)\right) P_{w,m,i}(z,t)+\frac{\partial P_{w,m,i}(z,t)}{\partial z}=0. \tag{4}$$

$$\left(\frac{\partial}{\partial t}+\gamma(x)\right) P_W(x,t)+\frac{\partial P_W(x,t)}{\partial x}=0. \tag{5}$$

$$\left(\frac{\partial}{\partial t}+\beta_i(u)\right) P_{r,m,i}(u,t)+\frac{\partial P_{r,m,i}(u,t)}{\partial u}=0, \qquad 1\leq m\leq k. \tag{6}$$

$$\left(\frac{\partial}{\partial t}+\delta(y)\right) P_R(y,t)+\frac{\partial P_R(y,t)}{\partial y}=0, \tag{7}$$

where

$$\lambda=\sum_{i=1}^{n}\lambda_i$$

and

$\alpha_i(z)\Delta \equiv$ the first order probability, that the ith component of class L_2 is commissioned for repairs in the interval $(z, z+\Delta)$, conditioned that it has been waiting up to time z.

$\beta_i(u)\Delta \equiv$ the first order probability that the ith component of class L_2 is repaired in the time interval $(u, u+\Delta)$ on the condition that it has not been repaired up to time u.

Similarly, we can attach meanings for $\gamma(x)\Delta$ and $\delta(y)\Delta$ for the components of class L_1. The relations between $\alpha_i(z)$, $\beta_i(u)$, $\gamma(x)$, and $\delta(y)$ and the corresponding probability densities $\Psi_i(z)$, $k_i(v)$, $S(x)$, and $\phi(y)$ are given by

$$S(x)=\gamma(x)\exp\left[-\int_0^x \gamma(x)dx\right]; \tag{8}$$

$$\phi(y)=\delta(y)\exp\left[-\int_0^y \delta(y)dy\right]; \tag{9}$$

$$\Psi_i(z)=\alpha_i(z)\exp\left[-\int_0^z \alpha_i(z)dz\right]; \tag{10}$$

$$k_i(u)=\beta_i(u)\exp\left[-\int_0^u \beta_i(u)du\right]. \tag{11}$$

Equations (1) through (7) are to be solved under the following boundary conditions:

$$P_{w,m,i}(o,t)=\lambda_i P_{o,m}(t); \tag{12}$$

$$P_W(o,t)=aP_{o,1}(t); \tag{13}$$

(14) $$P_{r,m,i}(o,t)=\int_0^\infty P_{w,m,i}(z,t)\,\alpha_i(z)\,dz;$$

(15) $$P_R(o,t)=\int_0^\infty P_W(x,t)\gamma(x)\,dx.$$

To solve Equations (1) through (7), we define the Laplace transform of the function $f(t)$ by $\bar{f}(s)$, given by the expression.

$$\bar{f}(s)=\int_0^\infty \exp(-st)f(t)\,dt \qquad Re(s)>0.$$

Let us assume that initially the system is in the operable state with all the K components of L_1 in working order so that,

$$P_{o,m}(O)=\begin{pmatrix}1 & \text{for } m=K\\ 0 & \text{for } m\neq K\end{pmatrix}$$

and

$$F_{F,m}(o)=P_{w,m,i}(o)=P_W(o)=P_{r,m,i}(o)=P_R(o)=0$$

$$(\text{for } i=1,2,\ .\ .\ .,n).$$

By using the initial conditions, taking the Laplace transforms of Equations (1) through (7) and (12) through (15), and integrating and simplifying them, we get

(16) $$\bar{P}_{w,m,i}(z,s)=\lambda_i \exp\left[-sz-\int_0^z \alpha_i(z)\,dz\right]\cdot \bar{P}_{o,m}(s);$$

(17) $$\bar{P}_{r,m,i}(u,s)=\bar{P}_{o,m}(s)\lambda_i\bar{\Psi}_i(s)\exp\left[-su-\int_0^u \beta_i(u)\,du\right];$$

(18) $$\bar{P}_W(x,s)=\bar{P}_{o,1}(s)\cdot a\cdot \exp\left[-sx=\int_0^x \gamma(x)\,dx\right];$$

(19) $$\bar{P}_R(y,s)=\bar{P}_{o,1}(s)a\cdot\bar{S}(s)\exp\left[-sy-\int_0^y \delta(y)\,dy\right];$$

(20) $$\bar{P}_{F,m}(s)=\bar{P}_{o,m+1}(s)\,\frac{a(1-b)}{s+\nu};$$

where $\bar{P}_{o,m}(s)$ is given by

(21) $$\left[s+\lambda+a-\sum_{i=1}^{n}\lambda_i\bar{\psi}_i(s)\bar{k}_i(s)\right]\bar{P}_{o,m}(s)=\left[ab+\frac{\nu a(1-b)}{s+\nu}\right]\bar{P}_{o,m+1}(s)\ 1\leq m\leq k-1$$

and

$$\left[s+\lambda+a-\sum_{i=1}^{n} \lambda_i \bar{\psi}_i(s)\bar{k}_i(s) \right] \bar{P}_{o,k}(s) = 1 + a\bar{S}(s)\bar{\phi}(s)\bar{P}_{o,1}(s). \tag{22}$$

Introduce the generating function defined by,

$$G(\theta, s) = \sum_{m=1}^{k} \bar{P}_{o,m}(s)\theta^m.$$

Multiplying (21) and (22) by appropriate powers of θ, we have,

$$G(\theta, s)\,[A - B/\theta] + \bar{P}_{o,1}(s)\,[B - a\bar{S}(s)\bar{\phi}(s)\theta^k] = \theta^k. \tag{23}$$

where,

$$A = s+\lambda+a-\sum_{i=1}^{n} \lambda_i \bar{\psi}_i(s)\bar{k}_i(s).$$

and

$$B = ab + \frac{\nu a(1-b)}{s+\nu}.$$

Setting $\theta = B/A$, we find

$$\bar{P}_{0,1}(s) = \frac{B^{k-1}}{C} \tag{24}$$

where,

$$C = A^k - a\bar{S}(s)\phi(s)B^{k-1}.$$

This gives

$$G(\theta, s) = \frac{\{1 + a\bar{S}(s)\bar{\phi}(s)B^{k-1}/C\}\theta^k - B^k/C}{A - B/\theta}.$$

Putting $m = 1$ in (21) and using (24), we have

$$\bar{P}_{o,2}(s) = \frac{A}{B} \cdot \frac{B^{k-1}}{C}.$$

Similarly

$$\bar{P}_{o,3}(s) = \frac{A^2}{B^2} \cdot \frac{B^{k-1}}{C}$$

and, in general,

$$\bar{P}_{o,m}(s)=\frac{A^{m-1}}{B^{m-1}}\cdot\frac{B^{k-1}}{C}.$$

and, in general,

$$\bar{P}_{o,m}(s)=\frac{A^{m-1}}{B^{m-1}}\cdot\frac{B^{k-1}}{C}.$$

Hence

$$\bar{P}_o(s)=\sum_{m=1}^{k}\bar{P}_{o,m}(s)=\frac{B^k-A^k}{C(B-A)}$$

where, $\bar{P}_o(s)$ is the Laplace transform of the probability that at time t, the system is in operable state irrespective of the number of working components in class L_1. Equation (20) gives

$$\bar{P}_{F,m}(s)=\frac{a(1-b)}{s+\nu}\cdot\frac{A^mB^{k-m-1}}{C}.$$

$\bar{P}_{w,m,i}(s)$, the Laplace transform of the probability that at time t the system is waiting for repairs due to failure of the ith component of L_2, is evidently given by,

$$\bar{P}_{w,m,i}(s)=\int_o^\infty \bar{P}_{w,m,i}(z,s)\,dz.$$

Integration and simplification of Equation (16) gives

$$\bar{P}_{w,m,i}(s)=\lambda_i\left[\frac{1-\bar{\psi}_i(s)}{s}\right]\bar{P}_{o,m}(s).$$

Similarly, the integration and simplification of Equations (17) through (19) gives

$$\bar{P}_{r,m,i}(s)=\lambda_i\bar{\psi}_i(s)\left[\frac{1-\bar{k}_i(s)}{s}\right]\cdot\bar{P}_{o,m}(s),$$

$$\bar{P}_W(s)=a\left[\frac{1-\bar{S}(s)}{s}\right]\bar{P}_{o,1}(s),$$

$$\bar{P}_R(s)=a\cdot\bar{S}(s)\left[\frac{1-\bar{\phi}(s)}{s}\right]\bar{P}_{o,1}(s),$$

where, $\bar{P}_{o,m}(s)$ has already been found and given by Equation (25). It may be verified that,

$$\bar{P}_W(s)+\bar{P}_R(s)+\sum_{m=1}^{k}\bar{P}_{o,m}(s)+\sum_{m=1}^{k-1}\bar{P}_{F,m}(s)+\sum_{m=1}^{k}\sum_{i=1}^{n}\left[\bar{P}_{w,m,i}(s)+\bar{P}_{r,m,i}(s)\right]=1/s.$$

PARTICULAR CASE

In the case when waiting after failure is exponential for both classes, L_1 and L_2, we will have to feed the new values for p.d.f. of $\bar{\psi}_i(s)$ and $\bar{S}(s)$ in the relation (25), as for exponential waiting time distribution we have,

$$\overline{\Psi}_i(s) = \int_o^\infty \nu_i e^{-\nu_i t} e^{-st} dt = \frac{\nu_i}{s+\nu_i}$$

and similarly

$$\bar{S}(s) = \frac{w}{s+w}.$$

So, we have

$$\bar{P}_{o,m}(s) = \frac{A^{m-1}}{B^{m-1}} \cdot \frac{B^{k-1}}{C},$$

Which agrees with Equation (38) as obtained by Parkash (5).

ACKNOWLEDGEMENTS

The author is extremely grateful to Dr. S. S. Srivastava, Chief Scientist and Dr. R. C. Garg, Senior Scientific Officer Gde I, for their keen interest and guidance in the preparation of this paper.

REFERENCES

[1] Cox, D. R., "The Analysis of Non-Markovion Stochastic Processes by Supplementary Variables," Proc. Com. 6 Phil. Soc *51*, 433–441 (1955).

[2] Flehinger, B. J., "Reliability Improvement through Redundancy at Various System Levels," I.B.M. J. Res. & Dev. *2*, 148 (1958).

[3] Halpensin, M., "Some Waiting Time Distributions for Redundant System with Repairs," Technometrics *6*, 27 (1964).

[4] Kulshrestha, D. K., "Dependability of a Standby Redundant Complex System with General Repair Time Distribution," OPSEARCH *3*, 139 (1966).

[5] Parkash, S. and Kumar, S., "Stochastic Behaviour of a Standby Redundant Complex System with Imperfect Switching over Device," Computing *5*, 136–143 (1970).

A DEFENSE ALLOCATION PROBLEM WITH DEVELOPMENT COSTS

Kenneth D. Shere

and

Edgar A. Cohen, Jr.

U.S. Naval Ordnance Laboratory, White Oak
Silver Spring, Maryland

ABSTRACT

In this paper, the mathematical model for the allocation of resources among a general mix of percentage vulnerable and of numerically vulnerable weapon systems is presented and solved. Percentage vulnerable systems consist of mobile weapons which are difficult to locate, but relatively easy to destroy once located; numerically vulnerable systems comprise easily located fixed base weapons which are difficult to destroy. The distinguishing feature of this analysis is the inclusion of development costs. The theory of max-min is extended as necessary to solve this problem. References are provided to a sequence of earlier versions of this problem.

I. INTRODUCTION

Strategic system decisions are subject to various levels of review. Technical, fiscal, and political factors are all introduced during the review process. Such problems as individual system performance and mutual support and protection between proposed and existing systems are also considered. Quantitative analysis of system mixes to support and justify selection decisions is beneficial, if not required. Such analysis should include the performance of individual systems taken individually, the performance of system mixes, costs to the enemy in defeating the various mixes, and protective interactions among the members of system mixes. In this paper a method for allocating resources among strategic weapon systems for deterrence purposes is presented.

In the following section, the mathematical model is presented. This model improves upon earlier models [1, 3, 4] by accounting for the price of admission. The price of admission includes development costs and we shall sometimes use these terms interchangeably; however, it specifically *excludes* research funds. In fact, before any system is considered for development it is assumed that preliminary research has been performed.

In sections III and IV, respectively, a method of solution is presented for an arbitrary mix of *percentage vulnerable* (PV) systems and for a general mix of PV and *numerically vulnerable* (NV) systems. The theory of max-min is extended in section III as necessary to solve this particular problem. The notation is summarized in the appendix.

II. MATHEMATICAL MODEL INCLUDING DEVELOPMENT COSTS

There are many ways in which system mix analyses can be conducted. In all of them the problem of characterizing the systems to be studied is of extreme importance. An intuitively appealing characterization of system alternatives results in a simple but highly versatile model. In this interpretation,

due to Dr. Thomas E. Phipps [3], defender retaliatory system candidates consist of two exclusive classes.

One class comprises systems which are difficult to locate but relatively easy to destroy once located. An example of such a mobile system is POLARIS. This type of weapon system is called *percentage vulnerable* because, for a fixed search effort by the attacker, a fixed percentage of the retaliatory weapons comes under attack. The other class consists of weapons that are easily located, but difficult to destroy. MINUTEMAN is an example of such a system. These systems are called *numerically vulnerable* because the attacker's effort must be distributed among all of the weapons of the system. We assume that each retaliatory weapon system can be attacked by only one weapon system of the y-player.

The retaliator allocates x_k resources, e.g., in billions of dollars, to the kth weapon system, which costs q_k to develop and $c_k n_k$ to procure; n_k is the number of weapons in the system and c_k is the procurement cost per weapon. The number of weapons in the kth system is

$$n_k = (x_k - q_k)/c_k.$$

If w_k is the throw-weight in megatons of a weapon in the kth system, the total throw-weight for the kth system is

$$w_k n_k = (w_k/c_k)(x_k - q_k).$$

Using the theory of random search [cf. 2], it can be shown [cf. 3, 4] that the fraction of weapons destroyed in the ith PV system is given by $1-\exp[-a_i(y_i - r_i)]$; y_i is the amount of resources the attacker allocates to blunting, on first strike, the corresponding retaliator's system and r_i is the attacker's development cost. a_i is the "vulnerability" of the ith system measured, e.g., in inverse billions of dollars. Consequently, the residual value of the ith percentage vulnerable system is

$$v_i(x_i - q_i) \exp[-a_i(y_i - r_i)], \tag{2.1}$$

where $v_i \equiv w_i/c_i$.

Throughout this article the subscript i will be used for percentage vulnerable systems and the subscript j will be used for numerically vulnerable systems.

For an NV system, the attacker's resources must be distributed among all the weapons of the system. Therefore, the survival probability is $\exp[-a_j(y_j - r_j)/(x_j - q_j)]$, and the residual value is

$$v_j(x_j - q_j) \exp[-a_j(y_j - r_j)/(x_j - q_j)]. \tag{2.2}$$

In deriving (2.1) and (2.2) it has been implicitly assumed that $x_k > q_k$ and $y_k > r_k$. Combining (2.1) and (2.2) and applying physical reasoning when the implicit assumptions are violated, we obtain the residual value for the retaliator's system mix:

(2.3a)

where

$$f_i(x_i, y_i) \equiv \begin{cases} 0 & : x_i \leq q_i \\ v_i(x_i - q_i) \exp\left[-a_i(y_i - r_i)\right] & : x_i > q_i,\ y_i > r_i \\ v_i(x_i - q_i) & : x_i > q_i,\ y_i \leq r_i \end{cases} \tag{2.3b}$$

and

$$f_j(x_j, y_j) \equiv \begin{cases} 0 & : x_j \leq q_j \\ v_j(x_j - q_j) \exp\left[-a_j(y_j - r_j)/(x_j - q_j)\right] & : x_j > q_j,\ y_j > r_j \\ v_j(x_j - q_j) & : x_j > q_j,\ y_j \leq r_j. \end{cases} \tag{2.3c}$$

The attacking y-player, having full knowledge of the retaliator's allocation, allocates his funds to minimize the x-player's retaliatory capability, $F(x, y)$. Consequently, the retaliator allocates his funds in a manner which maximizes this minimum; i.e., the objective is to determine the optimal strategies u for the retaliator and w for the attacker so that

$$V \equiv F(u, w) = \operatorname{Max}_x \left[\operatorname{Min}_y F(x, y)\right] \equiv \operatorname{Max}_x P(x) = P(u). \tag{2.4}$$

It is also desirable to know V. An unusually low value would indicate that a large infusion of funds by the defender is necessary.

Defining X and Y to be the retaliator's and attacker's total resources, respectively,

$$\sum_{i=1}^{n} x_i + \sum_{j=n+1}^{m} x_j = X \tag{2.5}$$

and

$$\sum_{i=1}^{n} y_i + \sum_{j=n+1}^{m} y_j = Y. \tag{2.6}$$

Of course, $x_k \geq 0$, $y_k \geq 0$, $a_k > 0$, $v_k > 0$, $r_k > 0$ and $q_k > 0$ for each weapon system. The limitations of this model and the procedures for determining the parameters a_k and v_k have been discussed by Phipps [4].

The mathematical model (2.3)–(2.6) with no development costs (i.e., $r_k = q_k = 0$ for all k) has been completely solved by Danskin [1]. The inclusion of development costs is one of several improvements in the existing theory necessary for a more realistic model.

For the present analysis the parameters c_k, v_k, q_k, a_k, and r_k are independent of x_k and y_k. The mathematical techniques are dependent upon the concavity with respect to x and convexity with respect to y of the payoff functions $F(x, y)$ in each of the regions defined in (2.3b) and (2.3c). If these parameters were allowed to vary such that $v_k(x_k)\ [x_k - q(x_k)]$ and $\exp\{-a(y_k)[y_k - r_k(y_k)]\}$ are respectively concave and convex, the model (2.3)–(2.6) becomes an example of the more general problem treated

in [5]. Unfortunately for the more general problem, the algorithm is more complicated and the computational time is considerably increased.

III. PV SYSTEMS

In this section the finite allocation problem with development costs is resolved for an arbitrary mix of PV systems, (2.3)–(2.6) with $m=n$. The case $m \geq n$ is discussed in section IV. It entails only trivial modifications of the procedure given in this section. The case $m=n$ is summarized by:

(3.1) Given: $$F(x, y)=\sum_{i=1}^{n} f_i(x_i, y_i);$$

(3.2) Constrained by: $$\sum_{i=1}^{n} x_i=X; \qquad \sum_{i=1}^{n} y_i=Y$$

$$x_i \geq 0; \qquad y_i \geq 0;$$

(3.3) Determine: u, w and V, where $V=F(u, w) \equiv \mathrm{Max}_x\, P(x) \equiv \mathrm{Max}_x\, [\mathrm{Min}_y\, F(x, y)]$.

We hypothesize that $X > \mathrm{Max}\ q_k$ and $Y > \mathrm{Max}\ r_k$. Physically this means that the retaliator considers only those systems for which he can afford to procure at least one weapon and that the attacker can afford at least a limited counter to any retaliatory system. Mathematically, these assumptions assure a positive residual value.

It is shown that (u, w) is the solution of a game defined on suitable subsets of the x-space and y-space. It is also shown that either $u_i=0$ or $u_i > q_i$, and a constructive procedure for solving (3.1)–(3.3) is then presented.

LEMMA 3.1: If $P(x)=F(x, \eta(x)) > 0$, then $x_k > q_k$ and $\eta_k > r_k$ for some k.

PROOF: Define $\Gamma \equiv \{i : x_{ii} > qq_i\}$ and let us suppose that $\eta_i \leq r_i$ for all $i \epsilon \Gamma$. Then $P(x)=\Sigma_\Gamma v_i(x_i - q_i)$. Select $k \epsilon \Gamma$ and define η^* by $\eta_i^*=0 (i \neq k)$ and $\eta_k^*=Y$. Since $Y > r_k$, $P(x) > F(x, \eta^*)$ contrary to the definition of $P(x)$. Consequently, there exists a $k \epsilon \Gamma$ such that $\eta_k > r_k$.

We note that $V=P(u) > 0$ because $P(x) > 0$ for $x=(X, 0, 0, \ldots, 0)$.

LEMMA 3.2: If $P(x)=F(x, \eta(x)) > 0$, $x_i \leq q_i$ implies $\eta_i=0$.

PROOF: Suppose, to the contrary, that $x_k \leq q_k$ and $\eta_k > 0$ for some $i=k$. Define $\Gamma=\{i : 0 \leq x_i \leq q_i\}$ and $\sigma \equiv \sum_{\Gamma} \eta_i$. From Lemma 3.1, $x_i > q_i$ and $\eta_i > r_i$ for some $i=k$. Define η^* by:

$$\eta_i^* \equiv \begin{cases} 0 & : i \epsilon \Gamma; \\ \eta_i & : i \epsilon \Gamma, \qquad i \neq k; \\ \eta_i + \sigma & : i=k. \end{cases}$$

$F(x, \eta^*) < F(x, \eta)$ contrary to the definition of $P(x)$ and η.

The following two lemmas are modifications of Gibbs' Lemma [cf. 1, p. 10]. The first modification is trivial, but is included to indicate the type of modification needed for the "committed resources problem" [5, 6]. The second modification allows the treatment of functions which are continuous, but not

differentiable with respect to x at $x_i = q_i$ nor with respect to y at $y_i = r_i$ for each i.

LEMMA 3.3: Let $f_i(x_i)$ be differentiable. Let $z = (z_1, \ldots, z_n)$ maximize $\sum_i f_i(x_i)$ constrained by

$$\sum_i x_i = X > 0; \qquad x_i \geq q_i \geq 0.$$

Then there exists a number λ so that

$$\begin{aligned} f_i'(z_i) &= \lambda : z_i > q_i; \\ &\leq \lambda ; z_i = q_i. \end{aligned}$$

LEMMA 3.4 (*Modified Gibbs' Lemma*): Let $f_i(x_i)$ be continuous with right- and left-derivatives. Let $z = (z_1, \ldots, z_n)$ maximize $\sum_i f_i(x_i)$ constrained by

$$\sum_i x_i = X, \qquad x_i \geq 0,$$

and $X > 0$. Then there exists a number λ such that

$$f_i'(z_i^+) \leq \lambda \qquad \text{for all } i;$$

$$f_i'(z_i^-) \geq \lambda \qquad \text{for all } i, z_i > 0.$$

Furthermore, if $f_i(x_i)$ is differentiable at $x_i = z_i > 0$, for some i, λ is unique.

PROOF: Suppose $z_i > 0$ and $0 \leq \epsilon \leq z_i$; define

$$F(\epsilon) \equiv f_i(z_i - \epsilon) + f_j(z_j + \epsilon) + \sum_{k \neq i,j} f_k(z_k).$$

The altered z still satisfies the constraints. Therefore, $F(\epsilon)$ is maximal at $\epsilon = 0$ and $F'(0) \leq 0$, i.e.

$$f_j'(z_j^+) \leq f_i'(z_i^-)$$

for all j. We now choose any λ such that

$$\operatorname{Max}_j f_j'(z_j^+) \leq \lambda \leq \operatorname*{Min}_{z_i > 0} f_i'(z_i^-).$$

We note that whenever $f_i(x_i)$ is differentiable at $z_i > 0$, for some i, the choice of λ is uniquely determined by $\lambda = f_i'(x_i)$.

The following two lemmas demonstrate that paying a portion of development costs without procuring any weapons is a waste of resources.

LEMMA 3.5: Either $w_i = 0$ or $w_i > r_i$.

PROOF: Define $A' \equiv \{i : 0 < w_i \leq r_i\}$. By Lemma 3.1, there exists $i = k$ such that $u_k > q_k$ and $w_k > r_k$. Select $i = k' \epsilon A'$ and define w^* as the n-dimensional vector:

$$w_i^* = \begin{cases} 0 & : i = k' \\ w_i & : i \neq k, k' \\ w_i + w_k & : i = k. \end{cases}$$

It follows that $F(u, w) > F(u, w^*)$ contrary to the definition of $\text{Max}_x \text{Min}_y$. Hence A' is empty.

LEMMA 3.6: Either $u_i = 0$ or $u_i > q_i$.

PROOF: Define $B \equiv \{i : u_i > q_i\}$ and $A \equiv \{i : w_i > r_i\}$. From Lemma 3.2, $A \subseteq B$. Then (3.1) becomes, at $x = u$,

$$F(u, y) = \sum_B f_i(u_i, y_i). \tag{3.4}$$

Application of the Modified Gibbs' Lemma to (3.4) yields

$$a_i v_i (u_i - q_i) \exp[-a_i(w_i - r_i)] = \mu : w_i > r_i; \quad 0 \leq \mu : w_i = 0 \tag{3.5}$$

For each $i \epsilon A$,

$$w_i = r_i + (1/a_i) \log [a_i v_i (u_i - q_i)/\mu]; \tag{3.6}$$

otherwise, $w_i = 0$. Substitution of (3.6) into the y-constraint of (3.2) yields

$$\sum_A \{r_i + (1/a_i) \log [a_i v_i (u_i - q_i)/\mu]\} = Y. \tag{3.7}$$

For each possible A, (3.7) either has a unique solution for μ or no solution. Unfortunately there is no *a priori* way of obtaining A, at this point of the analysis, even if u were known. The residual value of the retaliator's mix is:

$$V = \sum_A \mu/a_i + \sum_{B-A} v_i (u_i - q_i). \tag{3.8}$$

Thus A is determined by finding $\mu(A)$ for each possible choice of A and selecting the choice which minimizes V. We note that $a_i v_i (u_i - q_i) > \mu$ for each $i \epsilon A$, but the converse is not necessarily valid.

Assume that $0 < u_k < q_k$ for some k. We note that $k \notin B$ and by Lemma 3.2, $A \subseteq B$. Select $j \epsilon B$ and define u^* for suitable $\epsilon > 0$ by:

$$u_i^* = u_i : i \neq j, k;$$
$$u_j^* = u_j + \epsilon;$$
$$u_k^* = u_k - \epsilon.$$

We can choose ϵ so small that the possible choices for A are unaffected. Again applying the modified Gibbs' Lemma, (3.6)–(3.8) hold with u, w, μ, and V replaced by u^*, w^*, μ^*, and V^*. For each possible choice A one of the following statements is true:

$$\mu=\mu^* \text{ and } \sum_{B-A} v_i(u_i-q_i) < \sum_{B-A} v_i(u_i^*-q_i) : j\notin A;$$

$$\mu<\mu^* \text{ and } \sum_{B-A} v_i(u_i-q_i) = \sum_{B-A} v_i(u_i^*-q_i) : j\epsilon A.$$

Thus $V^*>V$, and $x=u$ could not have been the retaliator's optimal allocation. The lemma is established.

LEMMA 3.7: $P(x)\equiv \mathrm{Min}_y F(x,y)$ is a continuous function of x.

THEOREM 3.8: If the choice of A is unique (for a given u) and (u, w) solves (3.1)–(3.3), then (u, w) is also a solution of the game:

(3.9) Given: $$F(x,y)=\sum_A v_i(x_i-q_i)\exp[-a_i(y_i-r_i)]+\sum_{B-A} v_i(x_i-q_i)$$

(3.10) Constrained by: $$\sum_B x_i=X, \qquad \sum_A y_i=Y$$

$$x_i\geq q_i(i\epsilon B), \qquad y_i\geq r_i(i\epsilon A)$$

(3.11) Determine: $$V=\mathrm{Max}_\beta P_A(x)=\mathrm{Max}_\beta[\mathrm{Min}_\alpha F(x,y)]=\mathrm{Min}_\alpha[\mathrm{Max}_\beta F(x,y)]\equiv V_{B,A}$$

where

$$\beta(B)\equiv\{x : x_i=0 \text{ for } i\notin B \text{ and (3.10) satisfied}\};$$

$$\alpha(A)\equiv\{y : y_i=0 \text{ for } i\notin A \text{ and (3.10) satisfied}\}.$$

PROOF: Since A is unique and $P(x)=\mathrm{Min}\{P_{A'}(x) : A'\subseteq B\}$ is continuous at $x=u$, A is the minimizing set for each point x in some ϵ-neighborhood of $x=u$, $E\subseteq\beta$; that is, $P_A(x)=P(x)$. Hence $P_A(u)\geq P_A(x)$ for each $x\epsilon E$. Since A is unchanged for each $x\epsilon E$, $y(x)\epsilon\alpha$ and $F(x, y)$ is given by (3.9). From the concave-convex behavior of $F(x, y)$, (u, w) is a saddle-point. Since any saddle-point of (3.9) satisfying (3.10) is a solution to the game (3.9)–(3.11), the theorem is established.

Of course, (3.1)–(3.3) is not yet resolved because there is no *a priori* scheme for determining A, B and the uniqueness of A.

The uniqueness of A does not in itself present a real problem. Because of the political nature of this subject, none of the parameters are precisely known. For example, it is difficult for a retaliator to estimate his own resources or budget over a 5 to 10 year period; his estimate of the attacker's budget is even more tenuous. Consequently, a parameter analysis must be performed for any practical application to strategic systems. Except for isolated points the choice of A is generally unique. The isolated points may then be determined by continuity. Under these circumstances the attacker's choice is based partially on internal politics.

For the remainder of this section, A is assumed unique for some optimal allocation $x=u$. Under this hypothesis it is shown below that $A=B$ and an algorithm for determining B is presented.

Application of Gibbs' Lemma 3.3 to the game (3.9)–(3.11), which has an interior solution, yields for $i\epsilon B$:

$$v_i \exp\left[-a_i(w_i-r_i)\right]=\lambda_B : i\epsilon A; \tag{3.12a}$$

$$v_i \leqslant \lambda_B : i\notin A. \tag{3.12b}$$

Upon substitution of (3.12) into the y-constraint of (3.10), we determine λ from

$$\sum_A \left[r_i+(1/a_i)\log v_i/\lambda_B\right]=Y. \tag{3.13}$$

The residual value of (3.9)–(3.11) is

$$V_{B,A}=\lambda_B\left(X-\sum_A q_i-\sum_{B-A} q_i\right). \tag{3.14}$$

LEMMA 3.9: Let C be any set such that $B\supseteq C\supseteq A$ and let $V_{C,A}$ be the value of the game determined by (3.9)–(3.11) with B replaced by C. Then there is a number $\lambda_C\geqslant\lambda_B$ such that

$$V_{C,A}=\lambda_C\left[X-\sum_C q_i\right]. \tag{3.15}$$

PROOF: By applying Lemma 3.3 to (3.9)–(3.11) with B replaced by C, one finds for the optimal x, y and for $i\epsilon C$:

$$v_i=\lambda_C : x_i>q_i,\ y_i=0; \tag{3.16a}$$

$$v_i \exp\left[-a_i(y_i-r_i)\right]=\lambda_C : x_i>q_i,\ y_i>0; \tag{3.16b}$$

$$v_i\leqslant\lambda_C : x_i=q_i,\ y_i=0; \tag{3.16c}$$

$$v_i \exp\left[-a_i(y_i-r_i)\right]\leqslant\lambda_C : x_i=q_i,\ y_i>0. \tag{3.16d}$$

Hence

$$\sum_A \left[r_i+(1/a_i)\log(v_i/\lambda_C)\right]\leqslant\sum_A y_i=Y. \tag{3.17}$$

A comparison of (3.17) to (3.13) shows that $\lambda_C\geqslant\lambda_B$. From (3.16) the value is

$$V_{C,A}=\sum_{C-A} v_i(x_i-q_i)+\sum_A v_i(x_i-q_i)\exp\left[-a_i(y_i-r_i)\right]$$

$$= \lambda_C \left[X - \sum_C q_i \right].$$

THEOREM 3.10: $A = B$

PROOF: Suppose to the contrary that $A \subset B$. Set $C = A$ in Lemma 3.9 and compare (3.15) to (3.14). It is seen that $V_{A,A} > V_{B,A} = V$. Select $x' \epsilon \beta(A)$ such that $P_A(x') = V_{A,A}$. By Lemma 3.2,

$$P(x') = P_A(x') = V_{A,A} > V.$$

This contradicts the definition of V as $\text{Max}_x P(x)$. Hence $B = A$.

This theorem shows that the retaliator should not invest in a new system unless it is of sufficient value for the attacker to pay the penalty for at least a limited counter.

COROLLARY 3.11: *Let* $\mathscr{B} \equiv \{\Gamma \subseteq \{1, 2, \ldots, n\}$: (3.13) has a solution with $A = \Gamma\}$. Then $V = \text{Max}_{\mathscr{B}} V_{\Gamma,\Gamma}$

PROOF: By Theorems 3.8 and 3.10, $V = V_{A^*,A^*}$ for some $A^* \epsilon \mathscr{B}$. Suppose that

$$V_{A,A} \equiv \text{Max}_{\mathscr{B}} V_{\Gamma,\Gamma} > V.$$

Select $x_A \epsilon \beta(A)$ such that $P_A(x_A) = V_{A,A}$. From Lemma 3.2, $P(x_A) = P_A(x_A) > V$. This contradicts the definition of V.

Theorem 3.10 shows that $A = C = B$ in Lemma 3.9. It may be shown for any set D that

$$V_{D,D} = \lambda_D \left[X - \sum_D q_i \right],$$

where (3.16) holds with "C" replaced by "D". Consequently $\lambda_D \geq \lambda_D^*$, where λ_D^* is the solution of

$$\sum_D [r_i + (1/a_i) \log v_i/\lambda_D^*] = Y. \tag{3.13)'}$$

If $D = B$, $\lambda_D = \lambda_D^*$.

A procedure for solving (3.1)–(3.3) can be specified.

ALGORITHM 3.12:

1. Select $D \subseteq \{1, 2, \ldots, n\}$.
2. Solve (3.13)′ for λ_D^* and evaluate

$$V_D^* = \lambda_D^* \left(X - \sum_D q_i \right).$$

3. Go to step #1 until all possible choices of D are exhausted.
4. Determine B from $V_{B,B} = \text{Max}\, V_D = \text{Max}\, V_D^*$.

5. Solve for $w_i (i \in B)$ from (3.12a) and set $w_i = 0 (i \notin B)$.
6. Determine μ from (3.8), i.e., $\mu = V / \sum_B (1/a_i)$.
7. Determine $u_i (i \in B)$ from (3.5).
8. Set $u_i = 0 (i \notin B)$.

EXAMPLE: As an example we set $n = 2$. $r_1 = r_2 = 1/10$, $q_1 = 1/10$, $q_2 = 2/10$, $v_1 = 1$, $v_2 = 2/3$, $a_1 = 1$, $a_2 = 1/2$ and $X = Y = 1$. There are three possible choices for D, $\{1\}$, $\{2\}$, and $\{1, 2\}$. Considering these choices in turn (steps 1–3), we obtain Table 3.1.

TABLE 3.1. *Determination of Residual Value*

D	λ	V^*
1	0.40657	0.36591
2	0.425085	0.34007
1, 2	0.58380	0.40866

Continuing to follow Algorithm 3.12 one finds:

$V = 3\mu = 0.40866$

$B = \{1, 2\}$

$w = (0.637, 0.363)$

$\mu = V/3 = 0.13622$

$u = (1/3, 2/3)$

No procedure for obtaining the solution in closed form appears to be available. A few qualitative results, as

$$Y > \sum_B [r_i + (1/a_i) \log (v_i / v_{\text{Min}})],$$

where $v_{\text{Min}} \equiv \text{Min}_B v_i$, can be easily deduced and may help limit the number of cases that must be considered.

IV. GENERAL MIX OF PV AND NV SYSTEMS

In this section we first demonstrate that additional resources should be invested in at most one NV system. A method of solution for a general mix of PV and NV systems is then presented.

Lemmas 3.1–3.6 also apply to NV systems, as can be demonstrated with only trivial modifications in their proofs. As an equivalent to Theorem 3.8,

THEOREM 4.1: If the choice of A is unique and (u, w) solves (2.3)–(2.6) with $n = 0$, then (u, w) also is a solution of:

(4.1) Given: $$F(x, y) = \sum_A v_i (x_i - q_i) \exp [-a_i (y_i - r_i)/(x_i - q_i)] + \sum_{B-A} v_i (x_i - q_i)$$

(4.2) Constrained by: $$\sum_B x_i = X, \qquad \sum_A y_i = Y$$

$$x_i \geq q_i, \qquad y_i \geq r_i$$

(4.3) Determine: $$V = \text{Max}_B P_A(x) = \text{Max}_B \text{Min}_a F(x, y).$$

The proof is similar to the proof of Theorem 3.8. It is seen that $P_A(x) = P(x)$ in some neighborhood of $x = u$ and since $P_A(x)$ is a convex function, $P_A(u)$ is the maximum of $P_A(x)$ in α. We further note that $P_A(x)$ is a convex function on a convex set; hence, its maximum must occur at an extreme, or corner, point. We have, therefore, established a basic result.

THEOREM 4.2: At most one numerically vulnerable system should be developed.

Applying a trivial modification of Gibbs' Lemma to (4.1)–(4.3) yields

$$a_j v_j \exp\left[-a_j(y_j - r_j)/(x_j - q_j)\right] = \mu : y_j > r_j, \qquad j \in B. \tag{4.5}$$

We note that B has precisely one element, j. Applying (4.2) we determine μ from

$$r_j + [(x_j - q_j)/a_j] \log a_j v_j/\mu = Y.$$

By trying all possible combinations, B is selected so that $P(x)$ is maximized.

Finally we consider the general mix of PV and NV systems given by (2.3)–(2.6). As in the numerically vulnerable case it is determined [cf. 1, Theorem II, p. 64] that investment in NV systems should be limited to at most one weapon system. The amount of this investment is considered a parameter and we, thereby, reduce the problem to the PV problem considered in Section IV. Unfortunately, we have not yet determined an elegant way of finding the best value of the parameter; however, a value can be obtained by a computer search.

V. CONCLUSIONS

The problem of allocating resources to a general mix of percentage vulnerable and numerically vulnerable systems has been resolved. The price of admission has been included in the mathematical model. It has been shown that at most one numerically vulnerable system should receive additional resources. It has also been shown that a system should not be developed and procured by the retaliator unless it is of sufficient value to force the attacker to invest resources to develop and procure a counter.

The model is limited by the assumption of a one-to-one correspondence between attacking systems and retaliating systems. This means that none of the attacker's systems can attack two of the retaliator's systems and, conversely, none of the retaliator's systems are attacked by two attacking systems. The former case can be easily dispensed by renaming systems. The latter case introduces additional mathematical and philosophical problems. Suppose, for example, it is desired to include strategic bombers in the analysis. In this case the main problem is penetration rather than survivability. An airplane penetration model is more complex and may involve functions of the form $vx \exp[-a_1(y_1/x)^c - a_2 y_2]$. This function is, for suitable c, concave for some values of x and convex for other values of x for each fixed y. Thus, the mathematical difficulties are compounded. Philosophically it requires a marriage of penetration and survivability analyses.

Another limitation occurs in the classification itself. It would be desirable to investigate the effects of including a system which is intermediate between PV and NV. The extension of this model to include operating costs and time phasing of purchases is of major importance. Also of major importance is the extension of this model to include the value of committed resources.

ACKNOWLEDGMENT

The authors wish to thank Dr. T. E. Phipps, Jr. for discussions relating to this paper. This study

was supported by the Office of the Chief of Naval Operations, NOP–62, under task number NOL–459/ONR–50133.

Appendix

NOTATION

(a_k)	vulnerability of kth system
A	$\{i : w_i > r_i\}$
B	$\{i : u_i > q_i\}$
$\mathscr{B}$	$\{\Gamma \subseteq \{1, 2, \ldots, n\}$: (3.13) has a solution with A replaced by $\Gamma\}$
$f_k(x_k, y_k)$	residual value of retaliator's kth system (2.3)
$F(x,y)$	residual value of retaliator's system mix
i	refers to PV system
j	refers to NV system
m	total number of retaliator's (or attacker's) systems
n	number of PV systems
$P(x)$	security function, $\mathrm{Min}_y F(x, y)$
$P_A(x)$	$\mathrm{Min}_{\alpha(A)} F(x,y)$
q_k	retaliator's price of admission for kth system
r_k	attacker's price of admission for kth system
$u = (u_k)$	retaliator's optimal allocation
$v = (v_k)$	value (of the retaliator's kth system)
v_{Min}	$\mathrm{Min}_B v_i$
V	residual value of retaliator's system mix, $\mathrm{Max}_x P(x)$
$V_{B,A}$	value of game (3.9)–(3.11)
V_D^*	$\lambda_D^* \left(X - \sum_D q_i \right)$
$w = (w_k)$	attacker's optimal allocation
$x = (x_k)$	retaliator's allocation
X	retaliator's total resources
$y = (y_k)$	attacker's allocation
Y	attacker's total resources
$\alpha(\Gamma)$	$\{y : y_i = 0$ for $i \notin \Gamma$ and (3.10) is satisfied with $A = \Gamma\}$
$\beta(\Gamma)$	$\{x : x_i = 0$ for $i \notin \Gamma$ and (3.10) is satisfied with $B = \Gamma\}$
$\eta(x)$	optimal attacker's allocation for a given retaliator's allocation
λ, λ_B	constant of Gibbs' Lemma (a Lagrange multiplier)
λ^*_D	solution of equation (3.13)′
μ, μ_B	constant of Gibbs' Lemma (a Lagrange multiplier)

REFERENCES

[1] Danskin, J. M., *The Theory of Max-Min* (Springer-Verlag Inc., New York, 1967).

[2] Koopman, B. O., "The Theory of Search I," Operations Research 4, 324–346 (1956).

[3] Phipps, T. E., "Optimum Allocation of Effort for Deterrence," Operations Evaluation Group-Naval Warfare Analysis Group (currently a division of the Center for Naval Analysis, Arlington, Virginia 22209), OEG–NWG 43–60 (4 May 1960).

[4] Phipps, T. E., "Optimum Systems Choice for Strategic Retaliation: An Application of Max-Min Theory," NOL, White Oak, Silver Spring, Maryland 20910, Technical Report NOLTR 67–59, (1967).

[5] Shere, K. D., "A Resource Allocation Model with Development Costs and Prior Investments," NOL, White Oak, Silver Spring, Maryland 20910, Technical Report NOLTR 71–189 (1971).

[6] Shere, K. D. and E. A. Cohen, "Extension of Max-Min Theory to the Sunk Investment Problem Arising in Strategic Systems Selection Decisions," NOL, White Oak, Silver Spring, Maryland 20910, Technical Report NOLTR 71–59, (1971).

ON THE ISBELL AND MARLOW FIRE PROGRAMMING PROBLEM *

James G. Taylor†

Department of Operations Research and Administrative Sciences
Naval Postgraduate School
Monterey, California

ABSTRACT

A complete solution is derived to the Isbell and Marlow fire programming problem. The original work of Isbell and Marlow has been extended by determining the regions of the initial state space from which optimal paths lead to each of the terminal states of combat. The solution process has involved determining the domain of controllability for each of the terminal states of combat and the determination of dispersal surfaces. This solution process suggests a solution procedure applicable to a wider class of tactical allocation problems, terminal control attrition differential games. The structure of optimal target engagement policies in "fights to the finish" is discussed.

1. INTRODUCTION

An underdeveloped area [6] of the Lanchester theory of combat is target selection for combat among heterogeneous forces. This type of problem has been studied by Isbell and Marlow, who considered both a truncated stochastic (Lanchester) process by game theoretic means [12] and a terminal control (one-sided) differential game [11]. An attrition differential game is an idealized combat situation described by Lanchester-type equations [5], [18] over a period of time with choices of tactics available to both sides and subject to change with time. Terminal control attrition games only end when the course of combat has been steered to a prescribed state.

In developing a theory of target selection it is important to understand the dependence of allocation rules on the type of model chosen. Tactical allocation problems may be studied in two types of scenarios: (1) the prescribed duration battle and (2) the terminal control battle (a particular case of which is the "fight to the finish"). All the attrition examples in Isaacs' book [14] are of the first type (his "War of Attrition and Attack" is the continuous version of the tactical air war game [3], [4], [9] studied at RAND). Only Isbell and Marlow [11] and Weiss [19] have studied the terminal control problem. Unfortunately, Isbell and Marlow did not obtain a complete solution to their problem. They could not determine when certain terminal states of combat were reached. Weiss studied a problem which may be considered to be a generalization (two-sided version) of their problem. Writing many years before results were known beyond a small number of researchers, he did not use the subsequently well-known necessary conditions [1]. That his solution procedure [19] was heuristic is not surprising, since the simpler problem‡ [11] which he referenced in his paper had not been completely solved.

---

* Revision of part of a paper presented at the Spring Meeting of the Operations Research Society of America, San Diego, Western and Rocky Mountain Sections on March 20, 1970 in San Diego, California.

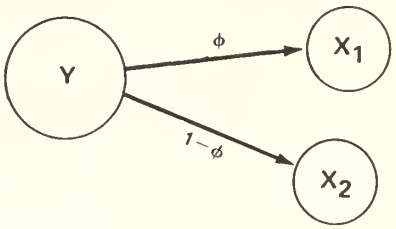
† This research was supported by the Office of Naval Research as part of the Foundation Research Program at the Naval Postgraduate School.

‡Since the original version of this paper, this problem has also been recently studied by S. Sternberg [17]. He employed a more geometric approach closely following that of R. Isaacs [14] and originally employed by Isbell and Marlow [11]. Sternberg obtained different results from those presented here.

Thus the obtaining of a complete solution to the Isbell and Marlow problem is viewed as a first step leading to the solution of more general tactical allocation problems. In this present note, we shall derive its solution by the Pontryagin maximum principle [16]. Generalizations of this procedure to differential games are indicated, and solution properties are discussed. In view of the close connection [1], [10] between optimal control and differential games (Isaacs), the terminology of these two fields is used somewhat interchangeably.

2. The Fire Distribution Problem

The situation considered by Isbell and Marlow [11] is the simplest problem of fire distribution: combat between an X-force at two force types (for example, riflemen and grenadiers) and a homogeneous Y-force (for example, riflemen only). This situation is shown diagrammatically below.



It is the objective of the Y-force commander to maximize his survivors at the end of battle and minimize those of his opponent (considering the utilities assigned survivors). This is accomplished through his choice of the fraction for fire, ϕ, directed at X_1. The battle terminates when one side or the other has been annihilated.

Mathematically the problem may be stated as

$$\underset{\phi(t)}{\text{maximize}}\ ry(T)-px_1(T)-qx_2(T) \text{ with } T \text{ unspecified}$$

$$\text{subject to: } \frac{dx_1}{dt}=-\phi a_1 y$$

$$\frac{dx_2}{dt}=-(1-\phi)a_2 y$$

$$\frac{dy}{dt}=-b_1x_1-b_2x_2 \tag{1}$$

$$x_1,x_2,y\geq 0 \qquad \text{and } 0\leq\phi\leq 1,$$

where all symbols are defined in the next section and with terminal states defined by (1) $x_1(T)=x_2(T)=0$ and (2) $y(T)=0$.

The terminal surface of the "realistic" (one-sided) game is seen to consist of five parts:

$C_1 : x_1(T)=0, \quad x_2(T)>0, \quad y(T)=0$

$C_2 : x_1(T)=0 \quad$ before $x_2(T)=0, \quad y(T)>0,$

$C_3 : x_1(T)=0 \quad$ after $x_2(T)=0, \quad y(T)>0,$

$C_4 : x_1(T)>0, \quad x_2(T)=0, \quad y(T)=0,$

$C_5 : x_1(T)>0, \quad x_2(T)>0, \quad y(T)=0.$

3. NOTATION

The symbols which are used in this paper are defined as follows:

$A=A(R, z)=[z^2(R-1)-R]/(z-1)^2,$

$B=B(R, z)=A(z-1)^2/z^2=[z^2(R-1)-R]/z^2,$

a_1, a_2, b_1, b_2 = constant attrition-rate coefficients,

C_i for $i=1, 2, 3, 4, 5$ = the ith part of the terminal surface as defined in section 2,

$D(C_i)$ = domain of controllability for C_i,

$g(P^0, R, z)$ = term in equation (57) of the locus of points for which $P_1=P_4$,

$h(P^0, R, z)$ = term in equation (58) for boundary surface between the regions from which optimal paths lead to C_1 and C_4,

p, q, r = utilities assigned to surviving X_1, X_2 and Y forces respectively,

$p_i(t)$ for $i=1, 2, 3$ = dual variable corresponding to $x_i(t)$ $(x_3(t)=y(t))$,

p_i^0 for $i=1, 2, 3$ = boundary condition for dual variable at $\tau=0$,

P_i for $i=1, 2, 3, 4, 5$ = payoff associated with an extremal leading to C_i,

$P^0=(x_1^0, x_2^0, y_0)$ = point in the initial state space,

$R=a_1b_1/(a_2b_2)$,

$s=s(x_1^0, x_2^0)=b_1x_1^0+b_2x_2^0,$

t_1 = time at which X_1 is annihilated, i.e., $x_1(t_1)=0$,

t_2 = first time at which $2b_1x_1(t_2)x_2^0+b_2(x_2^0)^2=a_2y^2(t_2)$ for an extremal leading to C_4,

T = total time for the battle,

$v=v(\tau)=a_2p_2(\tau)-a_1p_1(\tau),$

$V=V(x_1, x_2, y)$ = value of reduced game

$$w=\cosh\sqrt{a_2b_2}\tau_1(C_4)=\frac{a_1}{p_2^0}\frac{(b_1p_2^0+b_2p)}{(a_1b_1-a_2b_2)}$$

x_1, x_2, y = average force strengths; with initial values x_1^0, x_2^0, y_0,

$$z=\cosh\sqrt{a_2b_2}\tau_1(C_5)=\frac{a_1}{q}\frac{(b_1q-b_2p)}{(a_1b_1-a_2b_2)}=\frac{R-\delta}{R-1},$$

$\delta=a_1p/(a_2q)$,

ϕ = fraction of Y-fire directed at X_1,

τ = "backwards time" from the end of the reduced game, i.e., the time remaining before the end of the reduced game,

$\tau_1(C_i)$ = "backwards time" of the first switch in tactics for extremals leading to C_i,

$\hat{\tau}_1=T-t_1$ = length of time that Y fires at X_2 after X_1 has been annihilated for C_1 extremals.

4. SOLUTION PROCEDURE AND EXTENSIONS

Extremal paths (a path on which the necessary conditions for optimality are almost everywhere satisfied) may be obtained by routine application of Pontryagin's maximum principle [16] (the original authors used equivalent conditions independently developed by Isaacs [13]). However, in a terminal control problem we would like to know the domain of controllability [7] for each terminal state so that tactics are determined in terms of the initial conditions of combat (and also possibly time). We define the domain of controllability for a given terminal state to be that subset of the initial state space from which extremals lead to the terminal state. Furthermore, additional considerations (as outlined below) are required when the domains of controllability corresponding to different terminal states overlap.

The following procedure has been used to solve the above problem:

(a) Extremal control is determined by maximizing the Hamiltonian; since the state variables (force strengths) are nonnegative, the control depands, in many cases, only on relationships between the dual variables (marginal return from destroying target);

(b) From each separate terminal state, the time history of the dual variables is obtained by a backward integration of the adjoint system of differential equations; for a square law attrition process, the adjoint equations are independent of the state variables;

(c) For each terminal state the domain of controllability is determined by forward integration of the state equations using the time history of extremal control developed in (b); changes in control with time (existence of transition surface) may have to be considered in this step.

(d) The solution is determined at this point for regions of the initial state space which are covered by part of the domain of controllability for only one terminal state; one must also verify that the entire initial state space has been accounted for, since otherwise one may have overlooked some type of "singular" surface or even a terminal state;

(e) If domains of controllability overlap so that for a point of the initial state space contained in their intersection there is more than one extremal leading to the terminal surface, then one computes the payoff associated with each extremal; the optimal trajectory is selected from the extremals by comparing these values.

It is noted that Isbell and Marlow [11] stopped at step (b) above.

It seems appropriate to further discuss the situation when, for example, the domains of controllability corresponding to two different terminal states overlap. For points of the initial state space contained in the intersection of the two domains, there are two mutually exclusive possibilities: (1) for every point the payoff associated with the extremal leading to one terminal state always dominates that coresponding to the other, or (2) the payoff corresponding to extremals of one family does not always dominate that of the other. In the first case, the solution is simply determined: the optimal trajectories are the extremals corresponding to the more favorable payoff. In the second case, one must determine for what parts of the region of intersection each alternative is better. For the problem at hand, it will be seen that such a region of intersection may be divided into two subregions separated by a boundary surface. In each subregion the payoff corresponding to one family of extremals dominates that of the other. Isaacs (see [14], pp. 132–141) has referred to such a singular surface as a disperal surface, since when the geometry of the optimal trajectories is considered, optimal paths only lead away from a disperal surface. For the problem at and, it will be seen that for a certain range of model parameters, two dispersal surfaces are present in its solution.

TABLE I. *Solution to Target Selection Problem—Fight to Finish*

(Nonrestrictive assumption: $R > 1$, i.e., $a_1 b_1 > a_2 b_2$)

Terminal state	Optimal control	Domain of controllability
Case (1): $a_2 q \leq a_1 p$		
$C_1 \begin{cases} x_1(t_1) = 0 \\ x_2(T) > 0 \\ y(T) = 0 \end{cases}$	$\phi^*(t) = \begin{cases} 1 \text{ for } 0 \leq t \leq t_1 \\ 0 \text{ for } t_1 \leq t \leq T \end{cases}$	$a_1 b_1 y_0^2 < s^2 + (R-1)(b_2 x_2^0)^2$ $a_1 b_1 y_0^2 \geq s^2 - (b_2 x_2^0)^2$
$C_2 \begin{cases} x_1(t_1) = 0 \\ x_2(T) = 0 \\ y(T) > 0 \end{cases}$	$\phi^*(t) = \begin{cases} 1 \text{ for } 0 \leq t \leq t_1 \\ 0 \text{ for } t_1 \leq t \leq T \end{cases}$	$a_1 b_1 y_0^2 > s^2 + (R-1)(b_2 x_2^0)^2$
$C_5 \begin{cases} x_1(T) > 0 \\ x_2(T) > 0 \\ y(T) = 0 \end{cases}$	$\phi^*(t) = 1$ for $0 \leq t \leq T$	$a_1 b_1 y_0^2 < s^2 + (R-1)(b_2 x_2^0)^2$ $a_1 b_1 y_0^2 < s^2 - (b_2 x_2^0)^2$
Terminal state	Extremal control	Domain of controllability
Case (2): $a_2 q > a_1 p$		
C_1	Same as Case (1)	$a_1 b_1 y_0^2 < s^2 + (R-1)(b_2 x_2^0)^2$ $a_1 b_1 y_0^2 \geq s^2 + B(b_2 x_2^0)^2$
C_2	Same as Case (1)	Same as Case (1)
$C_4 \begin{cases} x_1(t_2) < 0 \\ x_2(T) = 0 \\ y(T) = 0 \end{cases}$	$\phi^*(t) = \begin{cases} 1 \text{ for } 0 \leq t \leq t_2 \\ 0 \text{ for } t_2 \leq t \leq T \end{cases}$	$a_1 b_1 y_0^2 < s^2 + (R-1)(b_2 x_2^0 T2$ $a_1 b_1 y_0^2 \geq R\{s^2 - (b_1 x_1^0)^2\}$ $a_1 b_1 y_0^2 \leq s^2 + R(R-1)(b_2 x_2^0)^2$
C_5	$\phi^*(t) = 0$ for $0 \leq t \leq T$	$a_1 b_1 y_0^2 < s^2 + (R-1)(b_2 x_2^0)^2$ $a_1 b_1 y_0^2 < R\{s^2 - (b_1 x_1^0)^2\}$ $a_1 b_1 y_0^2 \leq Rs^2 \left\{1 - \frac{1}{z^2}\right\}$
C_5	$\phi^*(t) = \begin{cases} 1 \text{ for } 0 \leq t \leq T - \tau_1 \\ 0 \text{ for } T - \tau_1 \leq t \leq \mathrm{T} \end{cases}$	$a_1 b_1 y_0^2 < s^2 + (R-1)(b_2 x_2^0)^2$ $a_1 b_1 y_0^2 < s^2 + B(b_2 x_2^0)^2$ $a_1 b_1 y_0^2 > s^2 + A(b_2 x_2^0)^2$ $a_1 b_1 y_0^2 > Rs^2 \left\{1 - \frac{1}{z^2}\right\}$

Definition of Times:

(a) t_1 is first t, such that $x_1(t_1) = 0$.

(b) t_2 is first t, such that $2b_1 x_1(t_2) x_2^0 + b_2 (x_2^0)^2 = a_2 y^2(t_2)$.

(c) τ_1 is determined by $\cosh \sqrt{a_2 b_2}\, \tau_1 = \dfrac{a_1}{q} \dfrac{(b_1 q - b_2 p)}{(a_1 b_1 - a_2 b_2)}$.

TABLE II. *Target Selection Problem—Fight to Finish*

Solution for $a_2q > a_1p$ and various values of parameters

(Nonrestrictive assumption: $R > 1$, i.e., $a_1b_1 > a_2b_2$)

Terminal state	Optimal control	Region of initial force levels
Case (a): $A \leq 0$, i.e., $R - \sqrt{R(R-1)} \leq \delta < 1$		
$C_1\begin{cases} x_1(t_1)=0 \\ x_2(T)>0 \\ y(T)=0 \end{cases}$	$\phi^*(t)=\begin{cases} 1 \text{ for } 0 \leq t \leq t_1 \\ 0 \text{ for } t_1 \leq t \leq T \end{cases}$	$a_1b_1y_0^2 < s^2+(R-1)(b_2x_2^0)^2$ $a_1b_1y_0^2 \geq s^2+B(b_2x_2^0)^2$
$C_2\begin{cases} x_1(t_1)=0 \\ x_2(T)=0 \\ y(T)>0 \end{cases}$	$\phi^*(t)=\begin{cases} 1 \text{ for } 0 \leq t \leq t_1 \\ 0 \text{ for } t_1 \leq t \leq T \end{cases}$	$a_1b_1y_0^2 > s^2+(R-1)(b_2x_2^0)^2$
$C_4\begin{cases} x_1(t_2)>0 \\ x_2(T)=0 \\ y(T)=0 \end{cases}$	$\phi^*(t)=\begin{cases} 1 \text{ for } 0 \leq t \leq t_2 \\ 0 \text{ for } t_2 \leq t \leq T \end{cases}$	$a_1b_1y_0^2 \geq R\{s^2-(b_1x_1^0)^2\}$ $a_1b_1y_0^2 \leq s^2+A(b_2x_2^0)^2$
$C_5\begin{cases} x_1(T)>0 \\ x_2(T)>0 \\ y(T)=0 \end{cases}$	$\phi^*(t)=\begin{cases} 1 \text{ for } 0 \leq t \leq T-\tau_1 \\ 0 \text{ for } T-\tau_1 \leq t \leq T \end{cases}$	$a_1b_1y_0^2 > s^2+A(b_2x_2^0)^2$ $a_1b_1y_0^2 < s^2+B(b_2x_2^0)^2$ $a_1b_1y_0^2 > Rs^2\{1-1/z^2\}$
C_5	$\phi^*(t)=0$ for $0 \leq t \leq T$	$a_1b_1y_0^2 \leq Rs^2\{1-1/z^2\}$ $a_1b_1y_0^2 < R\{s^2-(b_1x_1^0)^2\}$
Terminal state	Optimal control	Region of initial force levels
Case (b): $A > 0$, i.e., $0 \leq \delta < R-\sqrt{R(R-1)}$		
C_1	Same as Case (a)	$a_1b_1y_0^2 < s^2+(R-1)(b_2x_2^0)^2$ $a_1b_1y_0^2 \geq Rs^2-R\{b_1x_1^0[z^2(R-1)+R]+(2R)+b_2x_2^0\}z^2$ for $0 \leq x_1^0 < (b_2x_2^0)^2/(kb_1)$ $a_1b_1y_0^2 \geq h(P^0, R, z)$ for $0 \leq x_2^0 \leq kb_1x_1^0/b_2$
C_2	Same as Case (a)	Same as Case (a)
C_4	Same as Case (a)	$a_1b_1y_0^2 < s^2+(R-1)(b_2x_2^0)^2$ $a_1b_1y_0^2 \geq R\{s^2-(b_1x_1^0)^2\}$ $a_1b_1y_0^2 \leq h(P^0, R, z)$ for $0 \leq x_2^0 \leq kb_1x_1^0/b_2$
C_5	$\phi^*(t)=0$ for $0 \leq t \leq T$	$a_1b_1y_0^2 < s^2+(R-1)(b_2x_2^0)^2$ $a_1b_1y_0^2 > R\{s^2-(b_1x_1^0)^2\}$ $a_1b_1y_0^2 \leq Rs^2-R\{b_1x_1^0[z^2(R-1)+R]+(2R)+b_2x_2^0\}z^2$ for $0 \leq x_1^0 < (b_2x_2^0)^2/(kb_1)$

Explanation of Symbols:

(1) For t_1, t_2, and τ_1, see Table I.

(2) $k=[z^2-R(z-1)^2]/(2R)$.

The complete solution to this problem is shown in Tables I and II with supporting details presented in the next section. When $a_1p \geq a_2q$, the domains of controllability do not overlap so that the extremals are unique, and the extremal control turns out to be the optimal control. Hence, the solution procedure terminates at step (d) above. When $a_2q > a_1p$, the domains of controllability are shown in Table I as Case (2). For $R - \sqrt{R(R-1)} \leq \delta < 1$ where $\delta = (a_1p)/(a_2q)$, the domains of controllability don't overlap except for that corresponding to part C_4 of the terminal surface, which overlaps several others. In section 5.2 below it is shown how optimal trajectories are determined in the region of overlap and that extremals leading to C_4 can only be optimal for $a_1b_1y_0^2 \leq s^2 + A(b_2x_2^0)^2$. The resulting solution is shown in Table II as Case (a). It should be noted that nonbinding constraints, which define the regions in the initial state space for the various types of optimal control, have been deleted in constructing Case (a) of Table II from Case (2) of Table I. For $0 \leq \delta < R - \sqrt{R(R-1)}$, $D(C_1)$, the domain of controllability for the C_1 part of the terminal surface, and $D(C_5)$ have a nonempty intersection and similarly for $D(C_1)$ and $D(C_4)$. Hence, step (e) in our above solution procedure must be applied. In section 5.3 the determination of the two dispersal surfaces is shown. The resulting solution is shown in Table II as Case (b).

The above solution procedure may be easily extended to terminal control differential games (such as [19] in which the subsequently well-known necessary conditions [1] were not applied). Moreover, in two-sided problems (both the X-force and the Y-force are heterogeneous and each is faced with a fire distribution problem affected by the other's choice) this author has noted that not only may domains of controllability overlap with there being multiple extremals from a given point in the initial state space, but also the corresponding terminal states imply that the same player may either win or lose depending upon which extremal path is followed. This is illusory, however, since the player's opponent may block his steering the course of battle to a winning end point by use of a non-extremal strategy. Hence, the player must always lose.

We should note that in the existing theory of differential games [2 (a function independent of the X-force's strategy variable(s) plus a function independent of the Y-force's strategy variable(s))], [14] it is assumed that the optimal strategies are pure. Since the Hamiltonian is separable for two-sided fire distribution problems in the Lanchester theory of combat as formulated by Isbell and Marlow [11] and Weiss (see [18], pp. 94–95), it may be shown that there do exist pure strategy solutions to such problems [8]. Thus, this paper's developments are applicable to this class of problems.

5. DEVELOPMENT OF SOLUTION

The solution is actually derived for a "reduced" game (that portion of battle during which Y is faced with a choice problem). We illustrate here for extremals to C_1. It suffices to trace extremals up to t_1 when $x_1(t_1) = 0$, since $\phi^* = 0$ from then until the end of the game. The determination of the value, denoted by $V(x_1, x_2, y)$, of the reduced game, which is needed to determine the values of the adjoint variables on the terminal surface, and part of the solution originally obtained by Isbell and Marlow will not be repeated here although we shall outline the general steps.

The Hamiltonian is

$$H(t, x_i, p_i, \phi) = -\{p_1\phi a_1 y + p_2(1-\phi)a_2 y + p_3(b_1x_1 + b_2x_2)\}, \tag{2}$$

and the adjoint equations are

$$\frac{dp_1}{dt} = b_1p_3,$$

(3)
$$\frac{dp_2}{dt} = b_2 p_3,$$

$$\frac{dp_3}{dt} = p_1 a_1 \phi + p_2 (1-\phi) a_2,$$

with

(4)
$$p_1(t=t_1) = \text{unspecified}$$

$$p_2(t=t_1) = \frac{\partial V}{\partial x_2} = \frac{-q\sqrt{b_1 x_2}}{\sqrt{b_2 x_2^2 - a_2 y^2}}$$

$$p_3(t=t_1) = \frac{\partial V}{\partial y} = \frac{q a_2 y}{\sqrt{b_2}\,\sqrt{b_2 x_2^2 - a_2 y^2}}.$$

The extremal control ϕ^* is determined by the maximum principle, ie., $\underset{0 \le \phi \le 1}{\text{maximum}}\, H(t, x_i, p_i, \phi)$, and we also have that

(5)
$$H(t, x_i, p_i, \phi^*) = 0.$$

Equations (4) and (5) may be combined to yield that for extremals leading to C_1 we have

(6)
$$p_1(t=t_1^-) = \frac{-q a_2 \sqrt{b_2} x_2 (t=t_1)}{a_1 \sqrt{b_2 x_2^2 - a_2 y^2}},$$

or

(7)
$$p_1(t=t_1^-) = \frac{a_2}{a_1} p_2(t_1),$$

where

$$p_1(t=t_1^-) = \lim_{\substack{t \to t_1 \\ t < t_1}} p_1(t).$$

Obtaining a solution to this problem is simplified by the following considerations. Let $\tau = t_1 - t$ and define

(8)
$$v(\tau) = a_2 p_2(\tau) - a_1 p_1(\tau),$$

then we have

(9)
$$\frac{dv}{d\tau} = (a_1 b_1 - a_2 b_2) p_3(\tau),$$

with

(10)
$$v(\tau = 0) = a_2 p_2(\tau = 0) - a_1 p_1(\tau = 0),$$

and where (up until the first shift of tactics)

$$p_3(\tau) = p_3(\tau=0)\cosh\{\sqrt{\phi a_1 b_1 + (1-\phi)a_2 b_2}\,\tau\} - \frac{\phi a_1 p_1(\tau=0) + (1-\phi)a_2 p_2(\tau=0)}{\sqrt{\phi a_1 b_1 + (1-\phi)a_2 b_2}}\sinh\{\sqrt{\phi a_1 b_1 + (1-\phi)a_2 b_2}\,\tau\}. \tag{11}$$

It is easily seen that $p_3(\tau) > 0$ for all τ, and thus

$$\frac{dv}{d\tau} > 0 \qquad \text{for all } \tau, \tag{12}$$

when $a_1 b_1 > a_2 b_2$ by (9). As determined by the maximum principle and use of (2) and (8), the extremal control is given by

$$\phi^*(t) = \begin{cases} 1 \text{ for } v(\tau) > 0 \\ 0 \text{ for } v(\tau) < 0. \end{cases} \tag{13}$$

It is easy to show that it is impossible for $v(\tau)=0$ over any finite interval of time, and hence the possibility for any singular solution [15] to this problem is excluded. (The reader should note that the term "singular surface" is used in a slightly different sense by Isaacs (see [14], pp. 132–134 and 156–158) than in the control theory literature (see [10], p. 502, or [15], p. 4).)

By the symmetry of the problem it suffices to assume that $a_1 b_1 > a_2 b_2$, and then Equations (7), (8), (12), and (13) imply that extremals can actually lead to C_1. Furthermore, for $a_1 b_1 > a_2 b_2$ it is readily shown that no extremals lead to C_3 (i.e., $D(C_3)$ is void) and that extremals can only lead to C_4 when $x_2(t) > 0$ for $t < T$. We shall sketch the proof that no extremals for which $x_2(t)=0$ for $t < T$ can lead to C_4. By arguments similar to those used above for C_1 and using the value of the reduced game $V(x_1, x_2, y)$ and (2) and (5) with $\tau=0$ and $\phi^*=0$, it is readily shown that $v(\tau=0^+)=0$, where τ is the backwards time from the end of the reduce game, i.e., $x_2(\tau=0)=0$, and $v(\tau=0^+)$ is a one-sided limit through positive values. Considering (12), this implies that $v(\tau) > 0$ for $\tau > 0$, and hence by (13) X_2 is never fired at so that C_4 cannot be reached. The special case in which C_4 is reached but $x_2(t) > 0$ for $t < T$ is discussed in section 5.2.

In the next three subsections we present additional details of solution development: (1) determination of domains of controllability, (2) entry to C_4, and (3) determination of dispersal surfaces. This material represents the extension of the original work of Isbell and Marlow [11].

5.1. Determination of Domains of Controllability

One contribution of this note is to show how to determine the domains of controllability. There are two cases to consider.

CASE (1) $a_2 q \leq a_1 p$: These determinations are routine and consist in combining the time history of the extremal control, the nonnegativity requirements of the state variables, and the generalized square law

$$u^2(t_1) - u^2(t_2) = \{\phi a_1 b_1 + (1-\phi)a_2 b_2\}(y^2(t_1) - y^2(t_2)), \tag{14}$$

where $\phi(t) = \text{const.}$ in $t_1 \leq t \leq t_2$ and $u(t) = b_1 x_1(t) + b_2 x_2(t)$.

CASE (2) $a_2q > a_1p$: The condition for entry to C_2 is as before. There are two subcases for entry to C_5. First we consider the extremals for which

$$\phi^*(t) = \begin{cases} 1 \text{ for } 0 \leq t \leq T-\tau_1 \\ 0 \text{ for } T-\tau_1 \leq t \leq T. \end{cases} \tag{15}$$

Combining the above extremal control with the generalized square law (14), we obtain

$$R(b_2x_2(T))^2 + 2Rb_1b_2x_1(T-\tau_1)x_2(T) + (b_1x_1(T-\tau_1))^2 + 2(1-R)b_1b_2x_1(T-\tau_1)x_2^0 + a_1b_1y_0^2 - s^2 + (1-R)(b_2x_2^0)^2 = 0. \tag{16}$$

A backwards integration of the state equations of $0 \leq \tau \leq \tau_1$ with $x_2(\tau=0) = x_2(T)$ and $y(\tau=0) = 0$ yields that

$$b_2x_2^0 = b_2x_2(T)z + b_1x_1(T-\tau_1)(z-1), \tag{17}$$

since $x_2(t=T-\tau_1) = x_2^0$. It should be noted that in deriving (17) we have assumed that $x_1(T-\tau_1) > 0$. (16) and (17) may be combined to yield

$$(b_2x_2(T) - b_2x_2^0)^2 = (a_1b_1y_0^2 - s^2)/A, \tag{18}$$

and hence we require that the right-hand side of (18) be nonnegative in order that C_5 be reached. It may be shown that $A > 0$ leads to a contradiction to a result (32) derived below,* and hence $A \leq 0$ for extremals corresponding to (15) to lead to C_5. This corresponds to $R - \sqrt{R(R-1)} \leq \delta < 1$, since $A = [z^2(R-1) - R]/z^2$ and $z = (R-\delta)/(R-1)$. When $A \leq 0$, solution of (18) for $x_2(T)$ and the requirement that $x_2(T) > 0$ leads to

$$a_1b_1y_0^2 > s^2 + A(b_2x_2^0)^2. \tag{19}$$

Combination of the expression for $x_2(T)$ with (17) and the requirement that $x_1(T-\tau_1) > 0$ leads to

$$a_1b_1y_0^2 < s^2 + (b_2x_2^0)^2A(z-1)^2/z^2. \tag{20}$$

Next, we consider the extremals for which

$$\phi^*(t) = 0 \qquad \text{for } 0 \leq t \leq T. \tag{21}$$

It is readily shown that

$$y(t) = y_0 \cosh \sqrt{a_2b_2}t - \frac{(b_1x_1^0 + b_2x_2^0)}{\sqrt{a_2b_2}} \sinh \sqrt{a_2b_2}t, \tag{22}$$

so that the boundary, which occurs when $y(T) = 0$ for $T = \tau_1$, between this case and the previous one is easily seen to be given by

*Subsequent analysis has shown that externals corresponding to (15) do lead to C_5 when $A > 0$. No contradiction arises when $A > 0$, although points belonging to the (now nonempty) domain of controllability for extremals corresponding to (15), denoted as $D(C_5^g)$, do satisfy the inequality $a_1b_1y_0^2 > s^2 + B(b_2x_2^0)^2$. Hence, $D(C_5^g) \cap D(C_1)$ is nonempty for $0 \leq \delta < R - \sqrt{R(R-1)}$. Moreover, it has been shown that extremals corresponding to (15) which lead to C_5 are nonoptimal when $0 \leq \delta < R - \sqrt{R(R-1)}$.

$$y_0^2[\cosh \sqrt{a_2 b_2}\tau_1]^2 = \frac{(b_1 x_1^0 + b_2 x_2^0)^2}{a_2 b_2} \{[\cosh \sqrt{a_2 b_2}\tau_1]^2 - 1\}, \tag{23}$$

where the τ-time of the switch, determined from integration of the adjoint equations and the maximum principle, is given by

$$z = \cosh \sqrt{a_2 b_2}\tau_1 = \frac{a_1}{q} \frac{(b_1 q - b_2 p)}{(a_1 b_1 - a_2 b_2)}. \tag{24}$$

Noting that $\phi^* = 0$ for the entire battle when $T < \tau_1$ and rearranging, we obtain the result shown in Table I. Combination of the extremal control (21) with (14) also yields a quadratic equation for $b_2 x_2(T)$ which may be solved to yield

$$b_2 x_2(T) = -b_1 x_1^0 + \sqrt{s^2 - a_2 b_2 y_0^2}. \tag{25}$$

Thus, the payoff for extremals on which (21) holds is given by

$$P_5 = \frac{q}{b_2} \left\{ (b_1 x_1^0) \frac{(R-\delta)}{R} - \sqrt{s^2 - (a_1 b_1 y_0^2)/R} \right\}. \tag{26}$$

The requirement that $x_2(T) > 0$ leads to the result shown in Table I. When this latter condition or (19) is violated, then it is not possible for $x_2(T) > 0$ when $\phi^*(\tau) = 0$ for $0 \leq \tau \leq \tau_1$. Hence, we must investigate the possibility that $x_2(T) = 0$ (C_4 is entered), and this is done in the next section.

Finally, we consider extremals leading to C_1. As done above, we may combine the extremal control with (14) and the requirement that $x_2(T) > 0$ to obtain the first result shown in Table I. Similarly, $y(t_1) \geq 0$ implies that

$$a_1 b_1 y_0^2 \geq (b_1 x_1^0)^2 + 2 b_1 b_2 x_1^0 x_2^0. \tag{27}$$

Also, the payoff for extremals leading to C_1 is given by

$$P_1 = -\frac{q}{\sqrt{R}\, b_2} \sqrt{s^2 + (R-1)(b_2 x_2^0)^2 - a_1 b_1 y_0^2}. \tag{28}$$

It remains to develop conditions which distinguish between optimal paths leading to C_1 and C_5. (Thus, strictly speaking the result for C_1 shown in Table I Case (2) is for the region where extremals of $D(C_1)$ yield a larger payoff than extremals to C_5 when the two fields of extremals overlap.) We do this for $A < 0$, i.e., $z^2 < R/(R-1)$, since different arguments are required to determine optimal trajectories when $A > 0$ (see section 5.3). It will be shown that optimal paths can only lead to C_1 when the time remaining in the battle after destruction of the X_1 forces exceeds the maximum time to fire at X_2 for extremals reaching C_5.

Let $\hat{\tau}_1 = T - t_1$ denote the length of time that Y fires at X_2 after X_1 has been annihilated. Then optimal paths can only reach C_1 when $\hat{\tau}_1 \geq \tau_1(C_5) = \tau_1$, where $\tau_1(C_5)$ is the "backwards time" of the

switch in tactics for C_5 extremals. First, we shall show that extremals from P^0 with $x_1^0 > 0$ leading to C_5 yield a better payoff than those leading to C_1 for $a_1 b_1 y_0^2 \leq Rs^2\{1-1/z^2\}$ and $z^2 < R/(R-1)$. Thus, optimal paths can only lead to C_1 from initial points for which $a_1 b_1 y_0^2 > Rs^2\{1-1/z^2\}$. However, for such points the extremal strategy is to use $\phi^*(t=0)=1$ and to continue this until either the switching surface is encountered or X_1 is annihilated, i.e., $\phi^*(t)=1$ for $0 \leq t \leq$ minimum $(T-\tau_1, T-\hat{\tau}_1)$. Hence, optimal paths can only reach C_1 when $\hat{\tau} \geq \tau_1$.

Thus, it remains to sketch the proof that extremals from P^0 with $x_1^0 > 0$ leading to C_5 yield a larger payoff than those leading to C_1 for $a_1 b_1 y_0^2 \leq Rs^2\{1-1/z^2\}$ and $z^2 < R/(R-1)$. Setting $a_1 b_1 y_0^2 = Rs^2\{1-1/z^2\}$ in the expressions for P_5 and P_1 (Equations (26) and (28), respectively), it is easily shown that $P_5 > P_1$ for $z^2 < R/(R-1)$. Then since $\dfrac{\partial P_1}{\partial y_0} \geq \dfrac{\partial P_5}{\partial y_0} \geq 0$, it is readily seen that $P_5 > P_1$ for all y_0 such that $a_1 b_1 y_0^2 \leq Rs^2\{1-1/z^2\}$ when $A < 0$.

Combining the extremal control for C_1 with (14), we have that

$$y^2(t_1) = \frac{1}{a_1 b_1}\{a_1 b_1 y_0^2 - (b_1 x_1^0)^2 - 2b_1 b_2 x_1^0 x_2^0\}. \tag{29}$$

Also, since $\phi^*(t) = 0$ for $t_1 \leq t \leq T$, integration of the state equations (1) yields that

$$y(t) = y(t=t_1)\cosh\sqrt{a_2 b_2}(t-t_1) - x_2^0\sqrt{\frac{b_2}{a_2}}\sinh\sqrt{a_2 b_2}(t-t_1). \tag{30}$$

Since $y(t=T)=0$, we have that

$$\tanh\sqrt{a_2 b_2}\,\tau_1 = \frac{y(t=t_1)}{x_2^0}\sqrt{\frac{a_2}{b_2}}. \tag{31}$$

Since the hyperbolic cosine is a strictly increasing function of its argument, $\hat{\tau}_1 \geq \tau_1$ implies that

$$\cosh\sqrt{a_2 b_2}\,\hat{\tau}_1 \geq \cosh\sqrt{a_2 b_2}\,\tau_1 \geq 1.$$

Combination of this with (29) and (31), use of well-known identities for the hyperbolic functions, and some algebraic manipulation yields that

$$a_1 b_1 y_0^2 \geq (b_1 x_1^0 + b_2 x_2^0)^2 + \left\{\frac{z^2(R-1)-R}{z^2}\right\}(b_2 x_2^0)^2. \tag{32}$$

For $1 \leq z^2 \leq R/(R-1)$, it is easily seen that (32) implies (27) so that this latter condition need not be considered further.

5.2. Entry to Terminal State C_4

Terminal state C_4 is reached whenever X_2 can be annihilated by Y during the battle's last phase of length τ_1. The boundary conditions on the terminal surface are

$$\begin{aligned} &x_1(T) > 0 \qquad p_1(\tau=0) = -p \\ &x_2(T) = 0 \qquad p_2(\tau=0) = p_2^0, \text{ unspecified} \\ &y(T) = 0 \qquad p_3(\tau=0) = p_3^0, \text{ unspecified.} \end{aligned} \tag{33}$$

We also have that $\phi^*(t=T)=\phi^*(\tau=0)=0$, and hence by (33) and the transversality condition (5) it follows that $p_3^0=0$. It should be noted that p_2^0 remains unspecified on the terminal surface. It will be shown that p_2^0 is determined by conditions for extremals to reach C_4. Recalling the extremal control at $t=T$ and (13), we also must have that

$$a_1 p/a_2 < (-p_2^0). \tag{34}$$

Since $v(\tau=0)<0$ and $dv/d\tau>0$ by (12), there will be a switch in tactics in "backwards time" at $\tau=\tau_1(C_4)=\tau_1$. (In this section the notation τ_1 refers to extremals leading to C_4 when it is not denoted otherwise.) Thus, the extremal control is given by

$$\phi^*(t)=\begin{cases}1 \text{ for } 0\leq t\leq t_2\\ 0 \text{ for } t_2\leq t\leq T,\end{cases} \tag{35}$$

where $t_2=T-\tau_1$ and t_2 is such that

$$2b_1x_1(t_2)x_2^0+b_2(x_2^0)^2=a_2y^2(t_2). \tag{36}$$

From a backwards integration of the adjoint equations and the maximum principle, the time of the switch in tactics is given by

$$w=\cosh\sqrt{a_2b_2}\,\tau_1=\frac{a_1}{p_2^0}\frac{(b_1p_2^0+b_2p)}{(a_1b_1-a_2b_2)}=w(p_2^0). \tag{37}$$

We observe that from what region of the initial state space extremals lead to C_4 is determined by the value assigned to p_2^0. In other words, the value assigned to p_2^0 determines what initial force levels x_2^0 are to be totally destroyed. It seems reasonable to argue that $(-p_2^0)$ cannot exceed q, since otherwise we would implicitly be valuing X_2 survivors greater than stated for the problem. A more precise argument, of course, is to compare payoffs (i.e., P_4 and P_5, as done in section 5.1) for extremals leading to C_4 and C_5 from the same initial point P^0 corresponding to $p_2^0>q$, but this has not proven to be computationally tractable. Thus, recalling (34), we must have

$$a_1 p/a_2 < (-p_2^0) \leq q. \tag{38}$$

As done previously, we combine the extremal control (35) with (14) to obtain

$$(b_1x_1(t_2))^2-2(R-1)b_2x_2^0b_1x_1(t_2)+a_1b_1y_0^2-s^2-(R-1)(b_2x_2^0)^2=0, \tag{39}$$

which may be solved for $x_1(t_2)$ to yield

$$b_1x_1(t_2)=b_2x_2^0(R-1)+\sqrt{s^2+R(R-1)(b_2x_2^0)^2-a_1b_1y_0^2}, \tag{40}$$

whence the result shown in Table I by requiring that the quantity under the radical sign is nonnegative. The other entries for C_4 in Table I correspond to other parts of the terminal surface being impossible to reach.

It remains to show that optimal paths can only lead to C_4 from P^0 when

$$a_1 b_1 y_0^2 \leq s^2 + A(b_2 x_2^0)^2, \tag{41}$$

where $A = A(R, z)$ and $z = \cosh \sqrt{a_2 b_2}\, \tau_1(C_5)$. A backwards integration of the state equations (1) for $0 \leq \tau \leq \tau_1(C_4) = \tau_1$ with $x_1(\tau = 0) = x_1(t_2)$ and $x_2(\tau = 0) = 0$ yields that

$$x_1(t_2) = \frac{b_2 x_2^0}{b_1(w-1)}, \tag{42}$$

since $x_2(\tau = \tau_1) = x_2^0$ and $x_1(\tau = \tau_1) = x_1(t_2)$. Combining (39) and (42), we obtain

$$s^2 + A(R, w)(b_2 x_2^0)^2 - a_1 b_1 y_0^2 = 0. \tag{43}$$

For constant R, A may be considered to be a function of w. Considering the first two derivatives of $A(w)$ and $\lim_{w \to \infty} A(w)$, it is readily seen that for $1 \leq w \leq +\infty$, $A(w)$ has a global maximum at $w^* = R/(R-1)$ with $A(w^*) = R(R-1) > 0$. This latter result implies that for P^0 satisfying (41) the quantity under the radical sign in (40) is nonnegative (i.e., the corresponding entry in Table I is satisfied). It should be observed that $w(-p_2^0)$ is a strictly increasing function of $(-p_2^0)$ with

$$w\left(-p_2^0 = \frac{a_1 p}{a_2}\right) = 1 \text{ and } w(-p_2^0 = q) = z = \frac{R-\delta}{R-1} \leq \frac{R}{R-1}. \tag{44}$$

Since for $1 \leq w \leq w^*$, $A(w)$ is a strictly increasing function; then for $\frac{a_1 p}{a_2} \leq (-p_2^0) \leq q$, $A(w(-p_2^0))$ is a strictly increasing function whence (41). Also, combination of (37), (40), and (42) yields that

$$(-p_2^0(P^0)) = \frac{a_1 p\{b_2 x_2^0(R-1) + \sqrt{S(P^0, R)}\}}{a_2 \sqrt{S(P^0, R)}} \tag{45}$$

where

$$S(P^0, R) = s^2 + R(R-1)(b_2 x_2^0)^2 - a_1 b_1 y_0^2. \tag{46}$$

Equation (45) shows the dependence of dual variable values on the terminal surface upon the initial force levels P^0 of an extremal. It seems appropriate to observe that for $a_1 p$ [illegible] $\leq q$ we have that $0 \leq \tau_1(C_4) \leq \tau_1(C_5)$ and also $\tau_1(C_4) = \tau_1(P^0)$. The value implicitly used in (45) for X_2 survivors through p_2^0 is adjusted so that the X_2 forces are just annihilated at $t = T$. [illegible] note that by construction we have made $v(\tau = \tau_1(P^0)) = 0$ in our determination of $(-p_2^0(P^0))$ should make clear the reason that p_2^0 must be unspecified on C_4, for extremals corresponding to a given value of $x_1(T)$ and leading to C_4 there are an infinite number of different values for $\tau_1(C_4)$ extremals that result in this $x_1(T)$.

5.3. Determination of Dispersal Surfaces

It seems appropriate to note here some relationships between the inequalities which define the domains of controllability in Table I. It is always true that

$$s^2+B(b_2x_2^0)^2 \leq s^2+(R-1)(b_2x_2^0)^2. \tag{47}$$

For $R-\sqrt{R(R-1)} \leq \delta < 1$, the following hold

$$s^2+A(b_2x_2^0)^2 \leq s^2+(R-1)(b_2x_2^0)^2, \tag{48}$$

$$Rs^2\{1-1/z^2\} \leq s^2+(R-1)(b_2x_2^0)^2, \tag{49}$$

$$s^2+A(b_2x_2^0)^2 \leq s^2+B(b_2x_2^0)^2, \tag{50}$$

$$Rs^2\{1-1/z^2\} \leq s^2+B(b_2x_2^0)^2. \tag{51}$$

In (51), there is strict inequality for $x_1^0>0$ and $R-\sqrt{R(R-1)}<\delta<1$. For $0\leq\delta<R-\sqrt{R(R-1)}$, we also have

$$s^2+A(b_2x_2^0)^2 > s^2+B(b_2x_2^0)^2, \tag{52}$$

$$Rs^2\{1-1/z^2\} > s^2+B(b_2x_2^0)^2. \tag{53}$$

The proofs of the above are straight-forward and omitted.

When $0\leq\delta<R-\sqrt{R(R-1)}$, the above inequalities (52) and (53) imply that $D(C_1)$ and $D(C_5)$ for extremals on which $\phi^*(t)=0$ for $0\leq t\leq T$ overlap (i.e., there is a region in the initial state space from which extremals lead to both C_1 and C_5) and similarly for $D(C_1)$ and $D(C_4)$ when $a_1b_1y_0^2 \leq s^2+A(b_2x_2^0)^2$ in $D(C_4)$. Hence, we must consider step (e) of section 4's solution procedure.

First, we consider the determination of the dispersal surface which separates regions of the initial state space from which optimal paths lead to C_1 and C_5. Let us observe that for $z^2>R/(R-1)$ points exist for which

$$s^2+B(b_2x_2^0)^2 \leq a_1b_1y_0^2 \leq Rs^2\{1-1/z^2\}. \tag{54}$$

Setting $a_1b_1y_0^2=s^2+B(b_2x_2^0)^2$ in the expressions (26) and (28) for P_5 and P_1, respectively, it is easily shown that it is optimal to go to C_5 from such an initial point. The dispersal surface is the locus of points in the initial state space for which P_1 equals P_5 and extremals lead to both C_1 and C_5 on both sides of this surface. Equating (26) and (28), the locus of points for which $P_1=P_5$ is given by

$$a_1b_1y_0^2=Rx^2-R\{b_1x_1^0[z^2(R-1)+R]/(2R)+b_2x_2^0\}^2/z^2. \tag{55}$$

Furthermore, a dispersal surface actually exists, since it is readily shown that

$$Rs^2 - R\{b_1 x_1^0[z^2(R-1)+R]/(2R) + b_2 x_2^0\}^2/z^2 < Rs^2\{1-1/z^2\}, \tag{56}$$

when $z^2 > R/(R-1)$. Expression (56) and an above result mean that on both sides of the candidate dispersal surface extremals lead to both C_1 and C_5. Observing that $\frac{\partial P_1}{\partial y_0} > \frac{\partial P_5}{\partial y_0} > 0$ for $P^0 \epsilon D(C_1) \cap D(C_5)$, $x_1^0 > 0$, and $R > 1$, it is readily seen that $P_1 > P_5$ for all P^0 such that $a_1 b_1 y_0 > Rs^2 - R\{b_1 x_1 [z^2(R-1)+R]/(2R) + b_2 x_2^0\}^2/z^2$. Since we also must have $a_1 b_1 y_0^2 < R\{s^2 - (b_1 x_1^0)^2\}$ for extremals leading to C_5, the above dispersal surface only exists for $0 \leq x_1^0 < 2Rb_2 x_2^0/\{b_1[z^2 - R(z-1)^2]\}$.

Based on the above results there may also be a dispersal surface which separates optimal paths leading to C_1 and C_4 for $0 \leq x_2^0 \leq b_1 x_1^0[z^2 - R(z-1)^2]/(2Rb_2)$. Its determination is similar to the above, although the details are messier and haven't been completely worked out. Let

$$a_1 b_1 y_0^2 = g(P^0, R, z), \tag{57}$$

be the locus of points for which $P_1 = P_4$. Then, the boundary, which separates optimal paths leading to C_1 and C_4, is given by

$$a_1 b_1 y_0^2 = h(P^0, R, z), \tag{58}$$

where

$$h(P^0, R, z) = \begin{cases} s^2 + B(b_2 x_2^0)^2 & \text{for } g(P^0, R, z) < s^2 + B(b_2 x_2^0)^2 \\ g(P^0, R, z) & \text{for } s^2 + B(b_2 x_2^0)^2 \leq g(P^0, R, z) \leq s^2 + A(b_2 x_2^0)^2 \\ s^2 + A(b_2 x_2^0)^2 & \text{for } g(P^0, R, z) > s^2 + A(b_2 x_2^0)^2. \end{cases} \tag{59}$$

It should be noted that this boundary surface is a dispersal surface only when extremals to both C_1 and C_4 lie on both sides of $a_1 b_1 y_0^2 = g(P^0, R, Z)$.

6. STRUCTURE OF THE OPTIMAL ALLOCATION POLICIES

For square-law attrition it may be shown that the allocation of fraction of fire is always 0 or 1. In section 5 we have discussed the fact that the coefficient of the control variable ϕ in the Hamiltonian $H(t, x_i, p_i, \phi)$ cannot be equal to zero over any finite interval of time. This excludes the possibility of a singular control [15], and maximization of the Hamiltonian, which is a linear function of ϕ, yields that all fire is concentrated on one target type. This is not surprising, since our model assumes complete and instantaneous information [2] and that fire may be immediately shifted to a new target once the old one has been destroyed [5], [18].

With reference to Tables I and II, the condition that $a_1 b_1 > a_2 b_2$ may be interpreted to mean that there is more long range return for Y to engage X_1, i.e., more Y's will survive if this is done. Hence, when Y wins, he always engages X_1's while they are available. The condition $a_1 p < a_2 q$ means that at the end of battle there is greater payoff per unit time per Y soldier to engage X_2 not considering X_1's greater attrition effect against Y (short term gain at end of battle).

By the maximum principle and the well-known interpretation of the dual variables [1], Y always allocates his fire entirely to the target type yielding the greatest marginal return. However, marginal return evolves differently in winning or losing causes. When Y loses, he may switch from firing at X_1 entirely to firing at X_2 entirely to firing at X_2 entirely before the X_1 force has been annihilated. This happens when Y assigns utility to survivors of force type X_2 in excess of their kill rate against Y as compared to force type X_1, and X_1 is abundant enough not to be destroyed before the battle ends. Under these circumstances, whenever the X_2 forces are annihilated, this occurs at the very end of battle.

In this way, we see that tactics may depend on force levels. Moreover, the time at which the ranking of target priorities switches itself may depend upon force levels. We also see that Y's target priorities only switch with time in a losing case. This has occurred since a boundary condition at $t = T$ on one of the dual variables is dependent upon values of the state variables by a transversality condition. It may be shown that the structure of optimal allocation policies is different for the prescribed duration battle.

REFERENCES

[1] Berkovitz, L., "Necessary Conditions for Optimal Strategies in a Class of Differential Games and Control Problems," SIAM J. Control *5*, 1–24 (1967).

[2] Berkovítz, L., "A Survey of Differential Games," *Mathematical Theory of Control*, A. Balakrishnan and L. Neustadt (Ed.) (Academic Press, New York, 1967).

[3] Berkovitz, L. and M. Dresher, "A Game Theory Analysis of Tactical Air War," Operations Research *7*, 599–620 (1959).

[4] Berkovitz, L. and M. Dresher, "Allocation of Two Types of Aircraft in Tactical Air War: A Game Theoretic Analysis," Operations Research *8*, 694–706 (1960).

[5] Brackney, H., "The Dynamics of Military Combat," Operations Research *7*, 30–44 (1959).

[6] Dolansky, L., "Present State of the Lanchester Theory of Combat," Operations Research *12*, 344–358 (1964).

[7] Elgerd, O., *Control Systems Theory* (McGraw-Hill, New York, 1967).

[8] Friedman, A., "Existence of Value and of Saddle Points for Differential Games of Survival," J. Differential Equations *7*, 111–125 (1970).

[9] Fulkerson, D. and S. Johnson, "A Tactical Air Game," Operations Research *5*, 704–712 (1957).

[10] Ho, Y. C., "Review of the book *Differential Games by R. Isaacs*," IEEE Trans. on Automatic Control, *AC–10*, 501–503 (1965).

[11] Isbell, J. and W. Marlow, "Methods of Mathematical Tactics," Logistics Papers, No. 14, The George Washington University Logistics Research Project (Sept. 1956).

[12] Isbell, J. and W. Marlow, "Attrition Games," Nav. Res. Log. Quart. *3*, 71–94 (1956).

[13] Isaacs, R., "Differential Games—The Basic Principles of the Solution Process," The RAND Corporation, RM 1411, Vol. III (1954).

[14] Isaacs, R., *Differential Games* (John Wiley & Sons, Inc., New York, 1965).

[15] Johnson, C. D. and J. E. Gibson, "Singular Solutions in Problems of Optimal Control," IEEE Trans. on Automatic Control *AC–8*, 4–15 (1963).

[16] Pontryagin, L., Y. Boltyanski, R. Gamkrelidze, and E. Mishchenko, *The Mathematical Theory of Optimal Processes* (Interscience Publishers, Inc., New York, 1962).

[17] Sternberg, S., "Allocation Strategies," chap. 2 of append. D in "Development of Models for Defense Systems Planning," S. Bonder and R. Farrell (Eds.), Report No. SRL 2147 TR 70–2(U), Systems Research Laboratory, The University of Michigan (1970).

[18] Weiss, H., "Lanchester-Type Models of Warfare," *Proc. First International Conf. Operational Res.* (Oxford Press, 1957).

[19] Weiss, H., "Some Differential Games of Tactical Interest and the Value of a Supporting Weapon System," Operations Research 7, 180–196 (1959).

A NOTE ON *AN OPEN EXPANDING ECONOMY MODEL*

Oskar Morgenstern
New York University

and

Gerald L. Thompson
Carnegie-Mellon University

In volume 16 (December 1969) of the *Naval Research Logistics Quarterly*, we published a paper, "An Open Expanding Economy Model," in which the well-known von Neumann model, which up to then had been studied by many authors as a closed system, was transformed into the model of an open economy embedded in a "world economy." The consequences of this generalization were sufficiently important to cause immediate response from a number of correspondents. We also presented further work at the Conference on the von Neumann Model arranged by the Institute for Advanced Studies in Vienna, 1970 [5]. In 1971, the original paper [3] appeared with some changes in a French translation [4]. We would here like to record the changes from [3] to [4] for the convenience of readers of this *Quarterly*. We also add remarks on the work by Moeschlin and Rauhut [2] and on our own paper for the Warsaw "Symposium on Mathematical Methods of Economics" of 1972 [6].

After the appearance of [3] Professor J. Łos communicated to us some criticisms by Mrs. L. Mardon concerning the status of Assumption (A6) and its role in subsequent proofs. We acknowledge the correctness of these criticisms (but note that they do not invalidate any of our theorems), and change Assumption (A6) on page 449 of [3] to be the following:

$$\text{(A6)} \qquad t^{-}Ap^{+} > 0.$$

In other words, we change p^{-} in [3] to p^{+} here, as already noted in [4]. Then we change paragraphs 5 and 6 on page 449 of [3] to read as follows:

"The meaning of (A5)* will become clear later when we establish existence proofs. Briefly stated, however, its meaning is the following: *In order to have an economic solution to the open model which is essentially different from a solution to the closed model, then, even when it is operating at its minimum intensities, the economy must produce at least one product having a positive export price.* Stated this way, the requirement is economically significant, since, if this condition were not satisfied, the economy would either not export anything or else export only free goods. And if this were the case, the balance of payments condition would further require that only free goods be imported. Hence if the condition did not hold, the open economy would be, for all practical purposes, closed anyhow.

"Similarly the meaning of (A6) is the following: *when the economy evaluates its goods at the lowest possible prices, the export prices* p^{+}, *then even when operating at its minimum intensities the economy must have a positive demand to input at least one good which has a positive export price.* What this

*The Axiom (A5) was $t^{-}Bp^{+} > 0$.

assumption prevents is the possibility of operating at high intensity an industry that requires only inputs that are free, because they have a zero export price, and producing goods that have positive value. Stated in this way the condition is economically meaningful."

Because of these changes, further changes are required in the first paragraph on page 453 of [3]. The corrected paragraph should read as follows:

"Next consider the optimal solution when α is very large. Equation (12) becomes

$$(B-\alpha A)y-z^{+}+z^{-}=0. \tag{29}$$

Since $0 \leqq p^{+} \leqq y \leqq p^{-}$, we have

$$-z^{+}+z^{-}=-(B-\alpha A)y \geqq -Bp^{-}+\alpha Ap^{+}. \tag{30}$$

Multiplying on the left by x yields

$$-xz^{+}+xz^{-} \geqq -xBp^{-}+\alpha xAp^{+}. \tag{31}$$

Since $t^{-} \leqq x \leqq t^{+}$, we can set $x=t^{-}$ and deduce the inequality

$$-t^{-}z^{+}+t^{-}z^{-} \geqq -t^{-}Bp^{-}+\alpha(t^{-}Ap^{+}). \tag{32}$$

Now assumption (A6) says $t^{-}Ap^{+}>0$ so that we can make this quantity arbitrarily large by making α large. The objective function of the dual linear programs is $-t^{+}z^{+}+t^{-}z^{-}$ which differs by the term $-(t^{+}-t^{-})z^{+}$ from (32). Because the lefthand side of (29) decreases with α, z^{+} cannot increase with α, and we see that the objective function of the dual programs must also become arbitrarily large as α becomes large. This completes the proof."

The statement of Theorem 1 is unaffected by these changes and Theorem 2 was independent of (A6) at any rate. This raises the question of under what still more general assumptions our results are correct.

We repeat the remark given in [5] that R. L. Weil has provided in [7] a proof of the fact that the value of the linear programming problem discussed in the first paragraph of the proof of Lemma 2 on page 452 of [3] is a continuous function of the expansion parameter α.

In the first footnote on page 448 of [3] we noted that in any *basic* solution to the linear programming problem under consideration the conditions $w_i^{+}w_i^{-}=0$ and $z_j^{+}z_j^{-}=0$ always hold. In [2] Moeschlin and Rauhut properly remark (and in the spirit of our footnote) that nonbasic solutions *not* satisfying these properties may also exist, and that consequently, these conditions may if desired be added as additional assumptions to the model.

In a very interesting paper [1] A. Moeschlin has applied the model to show that by changing the control variables t^{+} and t^{-} it is possible to impose bounds on imports of goods, and that by changing import prices p^{-} it is possible to impose bounds on losses of industries.

Finally, several readers of [3] have asked how the control variables t^{+} and t^{-} (upper and lower bounds for production intensities) are to be fixed by the economy. This interesting question and its relation to long term planning is discussed in detail in our paper presented at the Warsaw symposium [6].

In view of the fact that our various studies of expanding economies extend over many areas and were published in a variety of journals, some of them of difficult access, we are currently engaged in writing a book which will cover in a unified way our previous work and in many cases make substantial extensions.

REFERENCES

[1] Moeschlin, O., "Eine Importbeschränkung im Morgenstern-Thompson (MT)—Aussenhandelsmodell," Zeitschrift für Nationalökonomie *31* (1971), pp. 425–442.

[2] Moeschlin, O. and B. Rauhut, "Eine Bemerkung zum Axiomensystem des Morgenstern-Thompson-Aussenhandelsmodel," Unternehmensforschung *4* (1971).

[3] Morgenstern, O. and G. L. Thompson, "An Open Expanding Economy Model," Naval Research Logistics Quarterly *16* (1969) pp. 443–457.

[4] Morgenstern, O. and G. L. Thompson, "Un modèle de croissance en économie ouverte," Économies et Sociétés *10*, 1703–1728 (1971).

[5] Morgenstern, O. and G. L. Thompson, "Further Consideration of 'An Open Expanding Economy Model,'" Zeitschrift für Nationalökonomie, Suppl. 1, pp. 23–30 (1971).

[6] Morgenstern, O. and G. L. Thompson, "Long Term Planning with Models of Static and Dynamic Open Expanding Economies," presented at the Symposium on Mathematical Methods of Economics, Warsaw, Poland, July 1972, Proceedings of the Symposium, J. Łos, Ed., North Holland Publishing Co. (to appear 1973).

[7] Weil, R. L., "Game Theory and Eigensystems," SIAM Review *X*, pp. 360–367 (1968).

A NOTE ON A PAPER BY HOUSTON AND HUFFMAN

W. E. Biles

Department of Aerospace and Mechanical Engineering
University of Notre Dame
Notre Dame, Indiana

and

J. W. Schmidt, Jr.

Department of Industrial Engineering and Operations Research
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

ABSTRACT

This paper calls attention to an experimental design which enhances the efficiency of the technique described by B. F. Houston and R. A. Huffman [3]. This alternative design is a modification of the simplex lattice design developed by Scheffé [6] for the mixture problem. While not possessing the optimal statistical properties of the central composite design proposed in [3], such as minimum bias and minimum variance, this modified design is optimum with respect to the number of design points employed; hence, it is the minimum cost design. Further, the design proposed in this note is excellent for use in multiple-block experimentation, which is very often required with constrained systems.

INTRODUCTION

Houston and Huffman [3] have presented a highly efficient and practical approach to experimenting with constrained systems. In essence, their technique consists of using designed experiments to develop models representing the real but unknown relationships between a set of response variables and the independent variables known to control the system under study, and then applying modified pattern search to solve for the optimum solution to the constrained optimization problem formed from these response models. The method presented in [3] appears to work very well for those systems for which an acceptable solution can be determined after a single block of experimentation and optimization. Biles [1] has encountered problems in which the solution determined after one block of experimentation and optimization was well removed from the true local optimum; with additional experimentation and optimization, however, progressively better solutions were observed.

This note describes an alternative experimental design which can be used in conjunction with the constrained pattern search procedure employed in [3] to yield a more economical approach to experimenting with constrained systems. This design uses the minimum number of design points commensurate with the postulated response model, provided the experimenter confines his attention to linear or quadratic models. Further, this design enables sequential experimentation in one or both of the following ways:

1. The experimenter can first examine a linear representation of the system under study, proceeding to a quadratic representation if the linear model proves inadequate. The design proposed here provides for the minimum number of design points for either the linear or the quadratic response model.

2. Given that the experimenter has examined a quadratic model in the ith block in a multiple-block experimental program, he can examine a structured subspace of the ith factor space in proceeding to the $(i+1)$st block. Further, he can reuse certain design points from the ith block.

EXPERIMENTAL DESIGNS

Retaining the notation of [3], let $X=(X_1, X_2, \ldots, X_k)$ represent a k-dimensional vector of independent variables. $F(X)$ and $C_j(X)$, $(j=1, 2, \ldots, m)$ are response functions which must be estimated through experimentation. The most common practice is to use either a linear response model given by

$$\eta=\beta_0+\sum_{i=1}^{k}\beta_i X_i, \tag{1}$$

or a quadratic response model given by

$$\eta=\beta_0+\sum_{i=1}^{k}\beta_i X_i+\sum_{i=1}^{k}\beta_{ii}X_i^2+\sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\beta_{ij}X_iX_j. \tag{2}$$

The linear model has $k+1$ terms and requires an experimental design with at least $k+1$ design points to estimate the $k+1$ coefficients. The quadratic model has $(k+1)(k+2)/2$ terms and requires at least $(k+1)(k+2)/2$ design points. The 2^k factorial design suggested in [3] as the linear structure in the central composite design has 2^k design points, which for large k is considerably more than needed to estimate the linear model. The central composite design has (2^k+2k+1) design points, which for large k exceeds the number required for the quadratic model. The 2^k factorial structure can be fractionated to yield more economical designs, but only at the sacrifice of some information about the model.

Scheffé [6] has developed a design for the mixture problem which has exactly the $(k+1)(k+2)/2$ design points needed to estimate the quadratic model. Although the experimental factor space for the mixture problem differs from that of the problem considered here, the concept of Scheffé's simplex lattice design can be extended to apply to this problem. Figure 1 illustrates the basic simplex design

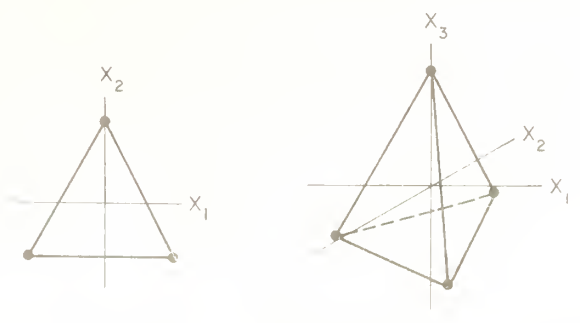


FIGURE 1. Simplex designs for estimating linear response models.

for two-variable and three-variable problems. The basic simplex design is a standard design for use in estimating linear models and contains $k+1$ design points in k-dimensional space, one point at each vertex of a k-dimensional regular simplex; for example, the simplex in two-dimensional space is an equilateral triangle. Myers [5] discusses the use of simplex designs for linear response models. A $\{k, 2\}$ simplex lattice in k-dimensional space is developed directly from the k-dimensional simplex. Figure 2 shows $\{k, 2\}$ simplex lattices for two and three-variable problems. Note that in addition to

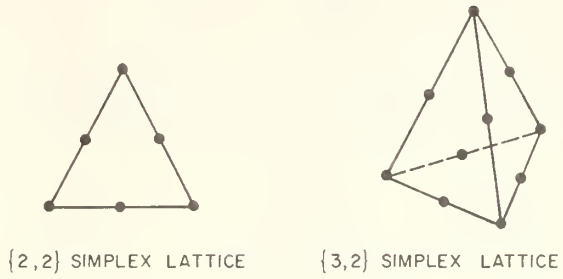


FIGURE 2. Simplex lattice designs for estimating quadratic response models.

the $k+1$ points at the vertices of the k-dimensional regular simplex, there is a design point at the midpoint of each edge of the simplex. This arrangement of points provides a symmetrical design in k-space employing exactly $(k+1)(k+2)/2$ points. For the problem considered here, the experimenter can fit the k-dimensional simplex inside the factor space formed by the known bounds,

$$a_i \leq X_i \leq b_i \qquad i = 1, 2, \ldots, k. \tag{3}$$

by placing the vertices of the simplex on the bounds (a_i, b_i). The orientation of the simplex within the factor space is left to the judgement of the experimenter. It is desirable to place the largest number of design points as close as possible to a suspected optimum. Hence, partial knowledge of the eventual solution enables the best orientation of the original simplex.

Figure 3 shows how the $\{k, 2\}$ simplex lattice design lends itself to sequential experimentation. In each successive block, a smaller simplex lattice is formed within the simplex lattice from the previous block. The subspace chosen is the one which lies closest to the computed optimum from the previous block. In this way k design points from the ith experimental block are reused in the $(i+1)$st block, so that only $k(k+1)/2$ new points are required to define the simplex lattice for the $(i+1)$st block.

A common practice with simplex designs is to augment the simplex or simplex lattice with a point at the centroid of the system. Such a design is called an "augmented" simplex or simplex lattice. While this center point is in excess of the minimum number of design points needed to estimate either the linear or quadratic model, it enhances the ability of the design to detect nonlinearity. The center point is often replicated to provide a means of testing the statistical significance of variation due to lack of fit.

Table I gives a comparison of the number of design points employed with factorial-based and simplex-based experimental designs in estimating linear and quadratic response models. The simplex and simplex lattice designs are consistently more economical than 2^k and central composite designs. This economy is achieved at the sacrifice of certain optimal properties of factorial and central composite designs, notably minimum bias and minimum variance. Myers [5] discusses these properties

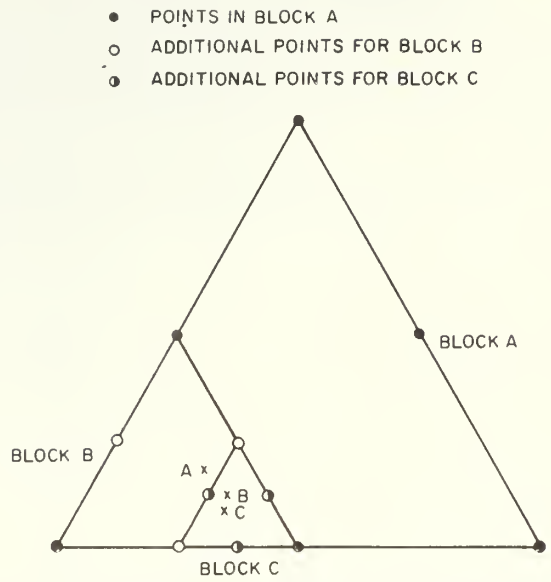


FIGURE 3. Sequential experimental blocks in a simplex lattice design.

in connection with central composite designs for estimating quadratic models. The bias and variance properties of simplex designs for estimating linear models are also discussed in [5]. The bias and variance properties of simplex lattice designs in estimating quadratic response models are not fully understood at present, but these designs have been shown to be very efficient when the effectiveness of the prediction result is stated in terms of cost. Biles [1] was able to locate a near-optimal set of conditions for a hypothetical solid propellant manufacturing process, using only two-thirds as many design points in a sequence of simplex lattice designs as would have been required with central composite designs. The following example illustrates the use of augmented simplex lattice designs in a hypothetical problem involving a constrained system.

TABLE I. *Number of Design Points Employed in Selected Response Surface Designs*

Model	Design	Number of factors (k)				
		2	3	4	5	6
Linear	2^k factorial	4	8	16	[a] 16	[a] 32
Linear	Augmented k-simplex	4	5	6	7	8
Terms in linear model		3	4	5	6	7
Quadratic	Central composite	9	15	25	[a] 27	[a] 45
Quadratic	Augmented $\{k, 2\}$ simple lattice	7	11	16	22	29
Terms in quadratic model		6	10	15	21	28

[a] $\frac{1}{2}$-fraction of a 2^k factorial design, resulting in loss of information on some interaction terms.

EXAMPLE PROBLEM

Consider the problem wherein one seeks to maximize the response $F(X)$, which is dependent upon two independent variables X_1 and X_2. The known bounds on the values that these independent variables may assume are given by

$$60 \leq X_1 \leq 92 \qquad 200 \leq X_2 \leq 280.$$

In addition to these bounds on the levels of the independent variables, there are specifications on two other system responses. These specifications are represented by the constraints

$$C_1(X) \geq 65$$

$$C_2(X) \geq 650.$$

Available information indicates that the best operating conditions for the system are likely to occur at higher values of X_1 and at middle values of X_2. This suggests that a simplex lattice design be oriented with the most points covering the higher values of X_1. Table II gives the design points and responses for the augmented simplex lattice design employed in the first experimental block.

TABLE II. *Design Points and Responses for First Block*

Design point	X_1	X_2	$F(X)$	$C_1(X)$	$C_2(X)$
1	60	280	2.00	70.0	500
2	92	280	2.50	45.0	450
3	76	200	2.03	53.5	650
4	68	240	2.27	72.0	695
5	76	280	2.30	61.0	560
6	84	240	2.33	62.0	660
7	76	253.3	2.65	70.0	675

TABLE III. *Regression Coefficients for First Block*

Term	$F(X)$	$C_1(X)$	$C_2(X)$
0	−5.84055	−456.83	−3717.7
1	0.02546	2.5691	42.908
2	0.05470	3.6614	25.049
11	−0.00056	−0.01471	−0.32255
22	−0.00015	−0.00679	−0.05718
12	0.00027	−0.00398	0.01629

NOTE: Number in term column denotes a subscript of coefficient b in quadratic response model.

Table III presents the regression coefficients describing quadratic response models for $F(X)$, $C_1(X)$, and $C_2(X)$. Using a constrained pattern search procedure by Moore, et al. [4], which is the Hooke and Jeeves method [2] modified to accommodate constraints, the indicated optimum solution is

$$X_1 = 80.7, \quad X_2 = 254.6, \quad F(X) = 2.318, \quad C_1(X) = 65.0, \quad C_2(X) = 650.$$

Hence, both $C_j(X)$ constraints are active in the indicated solution.

Table IV gives the design points and responses for the second block of experimentation, which is carried out within the simplex formed by points 4, 5, and 6 from the first block. The first block centroid (point 7) happens to be the centroid for the second block, so that only three new points are needed to describe a new augmented simplex lattice in the reduced region examined in block two. Design points 1, 2, and 3 in Table IV are the new points required for the second block.

Table V presents the regression coefficients for the quadratic models representing $F(X)$, $C_1(X)$, and $C_2(X)$ within the reduced region examined in block two. The optimum solution obtained from the pattern search is

$$X_1 = 78.3, \quad X_2 = 259.0, \quad F(X) = 2.905, \quad C_1(X) = 65.0, \quad C_2(X) = 650.$$

The known solution for this example problem is

$$X_1 = 78.7, \quad X_2 = 255.0, \quad F(X) = 2.715, \quad C_1(X) = 65.0, \quad C_2(X) = 650.$$

Hence, the indicated solution is within 0.5 percent of the true value for X_1 and within 1.6 percent for X_2. Moreover, the second block solution provides experimental confirmation of the solution obtained from the first block. It should be pointed out that, regardless of the experimental design employed, one must usually accept a solution in the neighborhood of, but not precisely equal to, the point giving the exact optimum.

TABLE IV. *Design Points and Responses for Second Block*

Design point	X_1	X_2	$F(X)$	$C_1(X)$	$C_2(X)$
1	72	260	2.48	80.0	635
2	80	260	2.80	62.5	620
3	76	240	2.40	71.5	850
4	68	240	2.27	72.0	695
5	76	280	2.30	61.0	560
6	84	240	2.33	62.0	660
7	76	253.3	2.65	70.0	675

TABLE V. *Regression Coefficients for Second Block*

Term	$F(X)$	$C_1(X)$	$C_2(X)$
0	−18.006	−1892.6	−3488.8
1	−0.26130	22.284	376.65
2	0.22492	9.380	−70.816
11	−0.00127	0.04166	−2.5768
22	−0.00071	−0.00831	0.1150
12	0.00190	−0.06922	0.00529

NOTE: Number in term column denotes subscript of coefficient b in quadratic model.

REFERENCES

[1] Biles, W. E., "A Constrained Optimization Methodology for Composite Materials Process Development," an unpublished Ph. D. dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia (1971).

[2] Hooke, R. and T. A. Jeeves, "Direct Search Solution of Numerical and Statistical Problems," J. Assoc. of Computing Machines *8*, 212 (1961).

[3] Houston, B. F. and R. A. Huffman, "A Technique Which Combines Modified Pattern Search With Composite Designs and Polynomial Constraints to Solve Constrained Optimization Problems," Nav. Res. Log. Quart. *18*, 91 (1971).

[4] Moore, C. F., C. L. Smith, and P. W. Murrill, "Multidimensional Optimization Using Pattern Search," IBM Share Library, LSU PATE SDA 3552 (1969).

[5] Myers, R. H., *Response Surface Methodology* (Allyn and Bacon, Rockleigh, N.J., 1971).

[6] Scheffé, H., "Experiments with Mixtures," J. Roy. Stat. Soc., Series B, *20*, 344 (1958).

☆U.S. GOVERNMENT PRINTING OFFICE: 1972-525-458/6

INFORMATION FOR CONTRIBUTORS

The NAVAL RESEARCH LOGISTICS QUARTERLY is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts and other items for publication should be sent to The Managing Editor, NAVAL RESEARCH LOGISTICS QUARTERLY, Office of Naval Research, Arlington, Va. 22217. Each manuscript which is considered to be suitable material for the QUARTERLY is sent to one or more referees.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Refereeing may be expedited if an extra copy of the manuscript is submitted with the original.

A short abstract (not over 400 words) should accompany each manuscript. This will appear at the head of the published paper in the QUARTERLY.

There is no authorization for compensation to authors for papers which have been accepted for publication. Authors will receive 250 reprints of their published papers.

Readers are invited to submit to the Managing Editor items of general interest in the field of logistics, for possible publication in the NEWS AND MEMORANDA or NOTES sections of the QUARTERLY.

NAVAL RESEARCH
LOGISTICS
QUARTERLY

SEPTEMBER 1972
VOL. 19, NO. 3
NAVSO P-1278

CONTENTS

ARTICLES Page

Mathematical Control Theory Solution of an Interactive Accounting Flows Model by Y. Ijiri and G. L. Thompson 411

Toward Optimal Bidding Strategies by A. C. Stedry and H. D. Brecht 423

Postoptimality Analysis in Zero-One Programming by Implicit Enumeration by G. M. Roodman 435

Statistical Auditing of Large-Scale Management Information Systems by S. E. Haber, F. W. Segel and H. Solomon 449

An Economic Model for Planning Strategic Mobility Posture by G. Bennington, L. Gsellman and S. Lubore 461

A Mixed Optimization Technique for the Generalized Machine Replacement Problem by S. P. Sethi and T. E. Morton 471

A Class of Inventory Models with Customer Impatience by M. J. M. Posner and B. Yansouni 483

Systems Approach to the Multistage Manufacturing Connected-Unit Situation by S. S. Ercan 493

On the Economic Application of Airlift to Product Distribution and Its Impact on Inventory Levels by D. Gross and A. Soriano 501

Bayesian Prior Distributions for Multi-Component Systems by J. B. Parker 509

Effect of Switch-Over Devices on Reliability of a Standby Complex System by P. Das 517

A Defense Allocation Problem with Development Costs by K. D. Shere and E. A. Cohen, Jr. 525

On the Isbell and Marlow Fire Programming Problem by J. G. Taylor 539

A Note on an Open Expanding Economy Model by O. Morgenstern and G. L. Thompson 557

A Note on a Paper by Houston and Huffman by W. E. Biles and J. W. Schmidt, Jr. 561

OFFICE OF NAVAL RESEARCH
Arlington, Va. 22217